

Discrete Beckmann Transport on Products of Simplices: Convergence, Sharp Commitment and a Schedule Rule

Johannes Herter

Harvard University

Abstract. Discrete Beckmann Transport Models (DBTM) generate a sequence in one step: a time-independent field on the probability simplex carries every point to a vertex, and the end point of its flow is learned from data through a conservation equation. Their theory covers one simplex per token and a linear schedule, the time at which a token is decided comes from a random-energy-model heuristic, and the schedule itself is not derived. We develop the theory of Beckmann transport on products of simplices with a separate schedule per block, which covers DBTM and *coarse-to-fine* codes that write a token as a tuple of vertices of small simplices. We compute the field in closed form; prove that for every block structure and every schedule the flow reaches the data in finite time from almost every start, with mean hitting time at most one, and that the one-step map pushes the prior to the data; prove a non-asymptotic two-sided commitment threshold and show that, in the signal-to-noise scale, every commitment has the same Gaussian profile; show that the flow decides the blocks in the order of their schedules for product-uniform data, and that the commitment thresholds survive correlation with a logarithmic shift; and bound the distance of a learned one-step sample from the data by its boundary error and conservation residual. In the time coordinate $\log(-\log(1-t))$ uniform training time is Gumbel distributed, which gives the schedule exponent that maximises the training time spent on the decision in closed form; it places commitment near $t = 1 - 1/e$ for every vocabulary, while the linear schedule of DBTM commits later as the vocabulary grows. In trained models on a synthetic language with 1,024 tokens this one-scalar change improves DBTM in every comparison we ran: with refinement-in-loop training, the valid sequences after eight evaluations rise from 51% to 72%. For the coarse-to-fine code we prove both its benefits and its limits: no extra signal at matched energy, but a state that fits through a narrow input layer, which the flat transport map provably does not. Trained at matched parameters, the coarse-to-fine model resolves the coarse level and uses context better than flat DBTM, but remains behind it after refinement; most of the loss comes from the block-level sampler, and a token-level commit rule largely closes the gap to DBTM’s linear schedule at four evaluations.

1 Introduction

Large language models generate text one token at a time [12, 49]. Each token requires a full evaluation of the network, conditioned on everything generated so far, so the time to produce a sequence grows linearly with its length, and the computation cannot be spread across positions. This has motivated a long line of work on generating many tokens at once, beginning with non-autoregressive translation [21] and continuing with diffusion models for discrete data. In these models all positions of a sequence start as noise and are refined jointly: masked and uniform discrete diffusion corrupt the tokens themselves [5, 29, 39, 42], while continuous approaches embed tokens in a vector space and transport a Gaussian to the data, as in image diffusion and flow matching [2, 22, 27, 44].

Parallel generation trades the sequential dependence on length for a sequence of denoising steps, and these models usually need tens to hundreds of network evaluations to match autoregressive quality. A second line of work therefore aims to compress the sampling process into a few steps or a single one, either by distilling a trained sampler [9, 45] or by training a few-step or one-step map directly [8, 17, 24, 35]. Beckmann

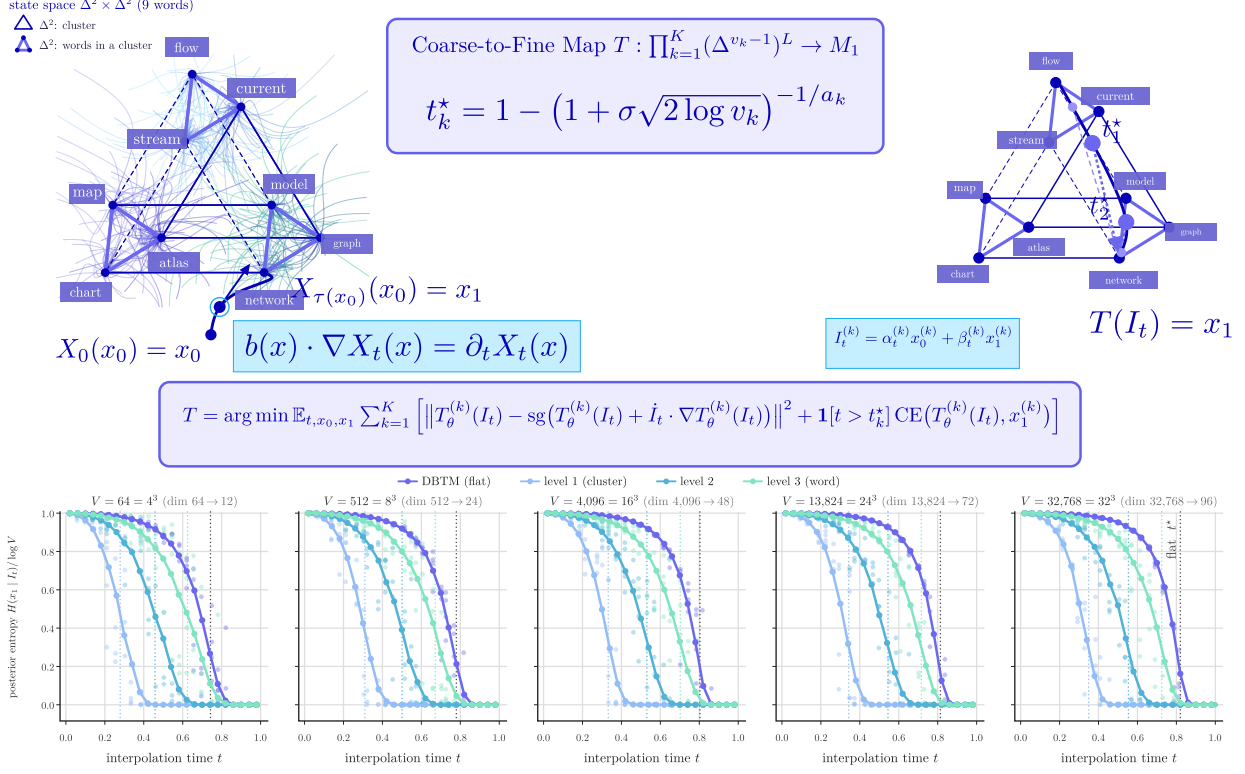


Figure 1 *Beckmann transport on a product of simplices.* Top left: the exact autonomous field (Proposition 3.3) on the coarse-to-fine state space $\Delta_{\text{cluster}}^2 \times \Delta_{\text{word}}^2$ of a nine-word vocabulary, the 3-3 duoprism, drawn through a linear map from its four-dimensional affine hull to $\mathbb{R}^3$; streamlines are first drawn into the word triangle of a cluster (violet) and only then resolve a word. Top right: the coarse-to-fine interpolant towards $x_1 = \text{network}$; it commits to the cluster at t_1^* and to the word at t_2^* (Theorem 4.5), and the transport map sends every point on it to x_1 . Bottom: Monte Carlo posterior entropy $H(x_1 | I_t) / \log V$ of a flat DBTM token (one schedule) and of the three levels of a coarse-to-fine code, with the commitment times of Theorem 4.5 as dotted lines. The flat curve is plotted at the same per-coordinate noise without matching the signal energy per token; Proposition 4.24 explains why this comparison favours the hierarchy.

transport belongs to the second group: it replaces the time-dependent velocity of flow matching by a single time-independent field and learns the end point of its trajectories from a conservation equation, without a teacher [25]. This paper is about one specific question in that setting, namely how the representation of a token, and the schedule with which each part of it is resolved, shapes the field, the one-step map and its training.

Continuous generative models on the probability simplex embed each token of a sequence as a vertex of Δ^{V-1} and transport a Gaussian prior to the vertex configurations of the data [14, 16, 24, 46]. Discrete Beckmann Transport Models (DBTM) [47] build on the autonomous flows of Beckmann transport [25]: they average the flow matching velocity over interpolation time, obtain a time-independent field $\mathbf{b}$, and learn the one-step map $\mathbf{T}$ that sends every point to the endpoint of its trajectory. The map solves the conservation equation $\mathbf{b} \cdot \nabla \mathbf{T} = 0$ with $\mathbf{T} = \text{id}$ on the data, and the residual of this equation can be minimised directly from data, without distilling a teacher.

Three parts of DBTM are not yet backed by theory. Its convergence theorem covers equal simplices and the linear schedule only. The time at which the interpolant commits to a vertex, which decides when the anchor loss is switched on, is derived from the random-energy model in the limit $V \rightarrow \infty$ [47, Thm. 4.1]. And the schedule is linear by choice; nothing says how much of the training distribution this spends on the decision that matters, which at language-model vocabularies ($V \approx 5 \cdot 10^4$) is a single late phase transition. A fourth point is structural: each position carries a V -dimensional state, read by a network through a layer much narrower than V . Language, by contrast, has coarse structure that can be decided before the exact

word [11, 31], and diffusion models are known to resolve hierarchical structure in stages [6, 40].

We develop the theory on products of simplices with a separate schedule per block. This covers flat DBTM, and it covers the simplest way to give DBTM this coarse structure: code each token as a tuple of vertices of small simplices and give every level its own schedule. Our contributions are:

1. **The field on products of simplices.** A hierarchical code places each position on $\prod_k \Delta^{v_k-1}$, with ambient dimension $\sum_k v_k$ instead of V (as small as $e \ln V$, Proposition 3.5). The field is explicit (Proposition 3.3); every block relaxes towards a point of its own simplex at the rate $a_k M(\mathbf{x})$, with one scalar speed M shared by all blocks (Lemma 3.6), and the product of simplices is forward invariant (Proposition 3.4).
2. **Convergence for every schedule.** For unequal blocks and every exponent $a_k > 0$ the flow is absorbed at the data in finite time for almost every start, with $\mathbb{E}\tau \leq 1$; the one-step map pushes the prior to the data and solves the conservation equation on an open set of full measure (Theorems 3.7 and 3.8). DBTM’s theorem covers equal simplices and the linear schedule only.
3. **Sharp commitment.** The posterior mass of the true vertex tends to one once the signal-to-noise ratio exceeds $\sqrt{2 \log(n-1)}$ by a margin and to zero below it, with explicit probabilities for every n (Theorem 4.2), which replaces the random-energy-model heuristic of DBTM [15, 47]. Levels commit at $t_1^* < \dots < t_K^*$ (Theorem 4.5), and with correlated levels and positions the thresholds shift by the logarithm of a bounded odds ratio (Section 4.5).
4. **The flow is coarse-to-fine.** The flow carries exactly the prior noise $\alpha_k(t)\mathbf{x}_0$ of the interpolant (Theorem 4.7), keeps the order of every block’s coordinates, so that the one-step map is the blockwise argmax for symmetric data (Theorem 4.9), and with exponent staggering resolves the coarse block of every position first (Corollary 4.13).
5. **Learning.** The DBTM loss splits over blocks and its fixed points solve the conservation equation (Proposition 5.1); the Wasserstein distance of the learned one-step sample from the data is at most the boundary error plus the conservation residual (Theorem 5.2 and Corollary 5.3); level-wise anchoring is Bayes-consistent after each level’s threshold (Proposition 5.5).
6. **A schedule rule for DBTM.** Every commitment has the same Gaussian profile in the signal-to-noise scale (Theorem 6.1). In the coordinate $\log(-\log(1-t))$ uniform training time is Gumbel distributed, the exponent of a schedule moves its commitment window and the prior scale sets its length (Proposition 6.2), and the exponent that maximises the training time spent on the decision is explicit (Theorem 6.3 and Corollary 6.4). It commits near $t = 1 - 1/e$ for every vocabulary. In training it improves DBTM in every comparison we ran (Table 2).
7. **Benefits and limits of a coarse-to-fine code.** At matched signal energy the finest level needs as much signal as a flat token (Proposition 4.24). Staggering the levels is an interval-packing problem with an explicit solution (Proposition 6.5). A linear input layer of rank r delays commitment to $\rho \approx s_n \sqrt{n/r}$ (Theorem 6.7) and hides the flat transport map, which puts a floor under the boundary error plus conservation residual of every flat model for product-uniform data; the coarse-to-fine state fits through the layer (Theorem 6.8 and Corollary 6.9).
8. **Experiments.** Exact computations check the threshold, the staggering and the fields (Section 7.1). Trained at matched parameters, the coarse-to-fine model resolves clusters and uses context better than flat DBTM but loses after refinement. Most of the loss comes from the block-level sampler; a token-level commit rule brings it level with flat DBTM under the linear schedule at four evaluations, while flat DBTM with a^* stays ahead (Section 7.2). At $V = 50,257$ the coarse-to-fine state makes a training step $2.5\times$ faster with $3\times$ less memory (Table 4).

The trained experiments use a synthetic language with known likelihood; Section 9 discusses what they do not show.

A running example. We illustrate the construction with a vocabulary of nine words in three clusters: $\{\textit{flow}, \textit{stream}, \textit{current}\}$, $\{\textit{map}, \textit{chart}, \textit{atlas}\}$ and $\{\textit{model}, \textit{network}, \textit{graph}\}$. Flat DBTM represents each word by a corner of the simplex Δ^8 , and the corners are equivalent: *stream* is as far from *current* as from *atlas*. A two-level code writes each word as a pair (cluster, word within the cluster), so that *network* becomes (3, 2),

and represents it by a corner of $\Delta^2 \times \Delta^2$, the 3-3 duoprism of Fig. 1. The first coordinate of the code carries the topic of the word and the second the word itself.

Roadmap. The results build on each other as follows. Convergence, the order of the levels and the minimiser of the training loss are all statements about the field $\mathbf{b}$, so Section 3 first computes it in closed form (Proposition 3.3). The one-step map is the end point of a trajectory and exists only if almost every trajectory reaches the data; DBTM proves this for equal simplices and the linear schedule, and Theorem 3.8 proves it for every block structure and every exponent. Section 4 then locates the time at which the interpolant commits to a symbol, for one simplex (Theorem 4.2) and level by level (Theorem 4.5), because these times fix when the anchor may be switched on and how the schedules should be chosen. Samples are produced by the flow and not by the interpolant, so Section 4.4 shows that the flow resolves the levels in the same order; and since in a real code the word depends on its cluster and neighbouring tokens depend on each other, Section 4.5 shows that the thresholds survive correlation with a logarithmic shift. Section 5 links the loss to the samples: Theorem 5.2 bounds the distance of a one-step sample from the data by the boundary error and the conservation residual, and the commitment times become rules for level-wise anchoring (Proposition 5.5) and for the share of training each level receives (Proposition 5.6). Section 6 derives the commitment profile and the schedule rule, and separates what the hierarchy can buy from what it cannot: every commitment has the same Gaussian profile (Theorem 6.1), uniform training time is Gumbel distributed in a logarithmic coordinate, which gives a closed-form schedule rule (Corollary 6.4), and a linear input layer of rank below V hides the flat transport map but not the coarse-to-fine one (Theorem 6.8). Section 7 checks the exact statements numerically, trains flat and coarse-to-fine models, and measures the cost of a training step. Complete proofs are in the appendix, and each opens with a short paragraph that states what has to be shown and why the chosen route fits.

2 Preliminaries

Sequences on the simplex. Let $\mathcal{V} = \{1, \dots, V\}$ be a vocabulary and $\mathbf{x} = (x^1, \dots, x^L) \in \mathcal{V}^L$ a sequence. Embedding token x^ℓ as the vertex $\mathbf{e}_{x^\ell}$ of $\Delta^{V-1} = \{\mathbf{p} \in \mathbb{R}_{\geq 0}^V : \mathbf{1}^\top \mathbf{p} = 1\}$ places a sequence on the vertex set $\mathcal{M}_1 = \{\mathbf{e}_1, \dots, \mathbf{e}_V\}^L \subset (\Delta^{V-1})^L$ of the product of simplices. The data distribution μ_1 is atomic on $\mathcal{M}_1$; the prior is $\mu_0 = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ on the ambient space $(\mathbb{R}^V)^L$.

Stochastic interpolants and the autonomous field. With $\mathbf{x}_0 \sim \mu_0$ and $\mathbf{x}_1 \sim \mu_1$ independent, the interpolant is $\mathbf{I}_t = \alpha(t)\mathbf{x}_0 + (1 - \alpha(t))\mathbf{x}_1$ with $\alpha(0) = 1$, $\alpha(1) = 0$ [2, 3]; DBTM uses $\alpha(t) = (1 - t)^a$, $a \geq 1$. The autonomous field of Beckmann transport [25, 47] averages the velocity over interpolation time:

$$\mathbf{b}(\mathbf{x}) := \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} [\dot{\mathbf{I}}_t \mid \mathbf{I}_t = \mathbf{x}], \quad t \sim \mathcal{U}[0, 1], \quad (1)$$

and defines the autonomous ODE $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$, $\mathbf{X}_0 = \mathbf{x}$. Averaging over t gives $\mathbf{b} = \int_0^1 \mu_t \mathbf{b}_t dt / \int_0^1 \mu_t dt$ and $\nabla \cdot \mathbf{j} = \mu_0 - \mu_1$ for the current $\mathbf{j} = \int_0^1 \mu_t \mathbf{b}_t dt$ [47, Prop. 3.1].

The difference from flow matching is that $\mathbf{b}$ has no time argument. Flow matching learns, for every t , the velocity of the interpolant at time t ; the autonomous field instead averages the velocities of all noise-data paths that pass through $\mathbf{x}$, over all times at which they do so. A sample is obtained by starting at a prior draw and following $\mathbf{b}$ until the trajectory stops, without a time schedule.

Convergence and the transport map. For the linear interpolant ($a = 1$) and any prior with a continuous, strictly positive density, the autonomous flow is absorbed at a vertex configuration in finite time $\tau(\mathbf{x})$ [47, Thm. 3.1]. The one-step map $\mathbf{T}(\mathbf{x}) := \lim_{s \rightarrow \tau(\mathbf{x})} \mathbf{X}_s(\mathbf{x})$ pushes μ_0 to μ_1 , is idempotent, and is the unique solution of the conservation equation

$$\mathbf{J}_{\mathbf{T}}(\mathbf{x}) \mathbf{b}(\mathbf{x}) = \mathbf{0} \quad (\mathbf{x} \notin \mathcal{M}_1), \quad \mathbf{T}(\mathbf{x}) = \mathbf{x} \quad (\mathbf{x} \in \mathcal{M}_1), \quad (2)$$

where $\mathbf{J}_{\mathbf{T}}(\mathbf{x}) \in \mathbb{R}^{D \times D}$ is the Jacobian matrix of $\mathbf{T}$ [47, Prop. 3.2].

Since two points on the same trajectory have the same end point, $\mathbf{T}$ is constant along trajectories, and its derivative in the direction of $\mathbf{b}$ vanishes; this is Eq. (2). The equation involves $\mathbf{T}$ and its Jacobian at a single point, so its residual can be evaluated on samples of the interpolant without integrating the flow, which is how DBTM trains without a teacher.

Training and refinement. DBTM regresses $\mathbf{T}_\theta$ onto a stop-gradient copy of itself advanced by an Euler step of Eq. (2), $\mathcal{L}_{\text{transport}} = \mathbb{E}[\|\mathbf{T}_\theta(\mathbf{I}_t) - \text{sg}(\mathbf{T}_\theta(\mathbf{I}_t) + \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t)\dot{\mathbf{I}}_t)\|^2]$, adds boundary, semigroup and anchor losses, and refines samples by committing confident positions and renoising the rest [47, Secs. 4–5]. Its anchor is switched on after the commitment time $t^* = 1 - (1 + \sigma\sqrt{2\log V})^{-1/a}$, derived from the random-energy model [15] in the limit $V \rightarrow \infty$ [47, Thm. 4.1].

In the flat construction all decisions about a token are made at t^* . Before it, the posterior of the endpoint is spread over many tokens, including tokens that a coarse description of the data would already rule out; after it, the token is fixed. A coarse-to-fine code, one of the cases covered by the next section, changes this behaviour.

3 Beckmann Transport on Products of Simplices

The construction changes only the space in which a token lives. A token becomes a tuple of corners of small simplices, one for each level of a code, and each level is given its own interpolation schedule. The field, the conservation equation and the loss are then defined exactly as in DBTM.

3.1 Hierarchical codes and per-level interpolants

Definition 3.1 (Hierarchical code). A *hierarchical code* of depth K is an injective map $\gamma : \mathcal{V} \rightarrow \prod_{k=1}^K [v_k]$, $w \mapsto (c^{(1)}(w), \dots, c^{(K)}(w))$, with $\prod_k v_k \geq V$. Level $k = 1$ is the coarsest. Token w at position ℓ is embedded as the stacked vector $(\mathbf{e}_{c^{(1)}(w)}, \dots, \mathbf{e}_{c^{(K)}(w)}) \in \prod_k \Delta^{v_k-1}$.

In the running example $K = 2$ and $v_1 = v_2 = 3$, so a token is a pair of one-hot vectors with six coordinates in total instead of nine. For $V = 50,257$, three levels of 37 symbols suffice, and a token has 111 coordinates. The results below hold for any injective code; how the code is obtained, for example by clustering token embeddings, does not enter the theory.

A state is $\mathbf{x} = (\mathbf{x}^{\ell,k})_{\ell \in [L], k \in [K]}$ with blocks $\mathbf{x}^{\ell,k} \in \mathbb{R}^{v_k}$, in the ambient space $\mathbb{R}^D$, $D = L \sum_k v_k$. A data configuration is a multi-index $\mathbf{j} = (j_{\ell,k}) \in \mathcal{J} := \prod_{\ell,k} [v_k]$ with vertex $\mathbf{e}_{\mathbf{j}} = (\mathbf{e}_{j_{\ell,k}})_{\ell,k} \in \mathbb{R}^D$, and the data manifold is $\mathcal{M}_1 = \{\mathbf{e}_{\mathbf{j}} : \mathbf{j} \in \mathcal{J}\}$. The data distribution is $\mu_1 = \sum_{\mathbf{j}} \pi_{\mathbf{j}} \delta_{\mathbf{e}_{\mathbf{j}}}$; configurations not in the image of γ get $\pi_{\mathbf{j}} = 0$.

Assumption 3.2 (Coarse-to-fine interpolant). Each level has a schedule $\alpha_k(t) = (1 - t)^{a_k}$ with $a_k > 0$ and a prior scale $\sigma_k > 0$. With $\mathbf{A}(t) = \text{diag}(\alpha_k(t)\mathbf{I}_{v_k})_{\ell,k}$ and $\mathbf{\Sigma} = \text{diag}(\sigma_k^2 \mathbf{I}_{v_k})_{\ell,k}$, let $t \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{x}_1 \sim \mu_1$ be independent and

$$\mathbf{I}_t = \mathbf{A}(t)\mathbf{x}_0 + (\mathbf{I} - \mathbf{A}(t))\mathbf{x}_1, \quad \text{i.e.} \quad \mathbf{I}_t^{\ell,k} = \alpha_k(t)\mathbf{x}_0^{\ell,k} + (1 - \alpha_k(t))\mathbf{x}_1^{\ell,k}. \quad (3)$$

The autonomous field $\mathbf{b}$ is Eq. (1) for this interpolant.

A level with $a_k > 1$ loses its prior noise early and is therefore decided early; with $a_k < 1$ it keeps its noise longer. A larger exponent for the coarse level, or a smaller prior scale σ_k with a common schedule, makes the cluster of a token identifiable before the word. The two options behave differently during training, as Proposition 5.6 shows.

Flat DBTM is the special case $K = 1$. Two ways to stagger the levels will matter: different exponents $a_1 > \dots > a_K$ with a common σ , and a common linear schedule $a_k = 1$ with different scales σ_k .

3.2 The exact autonomous field

Every later statement, from convergence to the minimiser of the training loss, is a statement about $\mathbf{b}$, so we need it in a form that can be analysed and evaluated. Given the time and the endpoint, the interpolant is

Gaussian with an explicit density, so the conditional expectation in Eq. (1) becomes an integral against the posterior of the pair.

For $t \in (0, 1)$ and $\mathbf{j} \in \mathcal{J}$ let $p(\mathbf{x} \mid t, \mathbf{j})$ be the density of $\mathbf{I}_t$ given t and $\mathbf{x}_1 = \mathbf{e}_j$,

$$p(\mathbf{x} \mid t, \mathbf{j}) = \prod_{\ell, k} (2\pi\alpha_k(t)^2\sigma_k^2)^{-v_k/2} \exp\left(-\frac{\|\mathbf{x}^{\ell, k} - (1 - \alpha_k(t))\mathbf{e}_{j_{\ell, k}}\|^2}{2\alpha_k(t)^2\sigma_k^2}\right), \quad (4)$$

and let $w_j(t, \mathbf{x}) := \pi_j p(\mathbf{x} \mid t, \mathbf{j})/q(\mathbf{x})$ with $q(\mathbf{x}) := \int_0^1 \sum_j \pi_j p(\mathbf{x} \mid s, \mathbf{j}) ds$ be the posterior density of $(t, \mathbf{x}_1 = \mathbf{e}_j)$ given $\mathbf{I}_t = \mathbf{x}$.

Proposition 3.3 (Exact field). *Under Assumption 3.2, for every $\mathbf{x} \notin \mathcal{M}_1$ the integrals below converge, and block (ℓ, k) of the autonomous field is*

$$\mathbf{b}^{\ell, k}(\mathbf{x}) = \int_0^1 \sum_{j \in \mathcal{J}} w_j(t, \mathbf{x}) \frac{a_k}{1-t} (\mathbf{e}_{j_{\ell, k}} - \mathbf{x}^{\ell, k}) dt. \quad (5)$$

The field is C^∞ on $\mathbb{R}^D \setminus \mathcal{M}_1$.

The weight $w_j(t, \mathbf{x})$ is the posterior over the time and the endpoint; the factor $a_k/(1-t) = -\dot{\alpha}_k/\alpha_k$ is the speed of level k . A level with a larger exponent a_k is pulled harder at every t ; this is the mechanism behind the coarse-to-fine streamlines of Fig. 1.

Each pair of a time t and a data configuration $\mathbf{j}$ that could have produced $\mathbf{x}$ contributes in proportion to its posterior weight, and pulls every block towards the corresponding corner. For small vocabularies the integral can be evaluated numerically to high accuracy, and all field plots in this paper are computed from Eq. (5) in this way rather than learned.

Before asking where trajectories end, we need to know where they can go. The convergence proof and the order results below rely on two facts about the geometry of $\mathbf{b}$: which corners attract a block, and whether the flow can leave the product of simplices.

Proposition 3.4 (Structure of the field). *For $\mathbf{x} \notin \mathcal{M}_1$ define the nonnegative coefficients $c_i^{\ell, k}(\mathbf{x}) := \int_0^1 \sum_{j: j_{\ell, k} = i} w_j(t, \mathbf{x}) \frac{a_k}{1-t} dt$, $i \in [v_k]$, and $C^{\ell, k}(\mathbf{x}) := \sum_i c_i^{\ell, k}(\mathbf{x})$. Then:*

- (i) Convex pull: $\mathbf{b}^{\ell, k}(\mathbf{x}) = \sum_{i=1}^{v_k} c_i^{\ell, k}(\mathbf{x}) (\mathbf{e}_i - \mathbf{x}^{\ell, k})$.
- (ii) Hull attraction: *along any solution of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$ in $\mathbb{R}^D \setminus \mathcal{M}_1$, the quantity $h^{\ell, k}(s) := \mathbf{1}^\top \mathbf{X}_s^{\ell, k} - 1$ satisfies $\dot{h}^{\ell, k} = -C^{\ell, k}(\mathbf{X}_s) h^{\ell, k}$; in particular $|h^{\ell, k}|$ is nonincreasing and the affine hull $\{\mathbf{1}^\top \mathbf{x}^{\ell, k} = 1 \ \forall \ell, k\}$ is invariant.*
- (iii) Forward invariance: *if $\mathbf{x} \in \mathcal{P} := \prod_{\ell, k} \Delta^{v_k-1}$, $\mathbf{x} \notin \mathcal{M}_1$ and $x_i^{\ell, k} = 0$, then $b_i^{\ell, k}(\mathbf{x}) = c_i^{\ell, k}(\mathbf{x}) \geq 0$. Hence $\mathbf{b}(\mathbf{x})$ lies in the tangent cone of $\mathcal{P}$ at $\mathbf{x}$, and every solution started in $\mathcal{P} \setminus \mathcal{M}_1$ stays in $\mathcal{P}$ while it exists.*

Part (i) says that a block is only attracted by corners of its own simplex. Part (ii) says that the flow returns to the affine hull of the simplices from any starting point, at a rate proportional to a_k , and part (iii) that it does not leave the product of simplices once it is there. The state can therefore be read as a tuple of probability vectors over the symbols of each level along the whole trajectory.

Proposition 3.5 (Ambient dimension). *A hierarchical code with $\prod_k v_k = V$ has ambient dimension $L \sum_k v_k \geq LKV^{1/K}$ per sequence, with equality iff $v_k = V^{1/K}$ for all k . For $V \geq 2$, over real $K > 0$, $KV^{1/K}$ is minimised at $K = \ln V$ with value $e \ln V$.*

For GPT-2's vocabulary ($V = 50,257$) the bound is $e \ln V \approx 29.4$ coordinates per position; the integer code $K = 3$, $v_k = 37$ ($37^3 = 50,653 \geq V$) uses 111, against 50,257 for flat DBTM.

The input and output layers of the map, and the Jacobian–vector product in the transport loss, scale with this number; Table 4 measures the effect on the cost of a training step.

3.3 Convergence for every schedule

The one-step map is the end point of a trajectory, so it is only defined if trajectories end. DBTM proves that they do for one simplex per token and the linear schedule. Neither the restriction to equal blocks nor the restriction to $a = 1$ is innocuous here: for $a_k > 1$ the schedule α_k has vanishing derivative at $t = 1$, and a level could in principle approach its corner too slowly to reach it. We first rewrite the field in a form that separates a common speed from the level rates.

Write $\mathcal{M}_1^+ := \{e_j : \pi_j > 0\} = \text{supp } \mu_1$ and $\Omega := \mathbb{R}^D \setminus \mathcal{M}_1^+$. For $\mathbf{x} \in \Omega$ let

$$M(\mathbf{x}) := \int_0^1 \sum_j \frac{w_j(t, \mathbf{x})}{1-t} dt, \quad \bar{\mathbf{p}}^{\ell,k}(\mathbf{x}) := \frac{1}{M(\mathbf{x})} \int_0^1 \sum_j \frac{w_j(t, \mathbf{x})}{1-t} e_{j_{\ell,k}} dt. \quad (6)$$

M is a posterior average of the inverse remaining time $1/(1-t)$, and $\bar{\mathbf{p}}^{\ell,k}(\mathbf{x}) \in \Delta^{v_k-1}$ is the posterior mean of the level- k symbol at position ℓ under the same tilted posterior.

Lemma 3.6 (Common speed). *The field of Proposition 3.3 is C^∞ on Ω , $1 \leq M(\mathbf{x}) < \infty$ there, and*

$$\mathbf{b}^{\ell,k}(\mathbf{x}) = a_k M(\mathbf{x}) (\bar{\mathbf{p}}^{\ell,k}(\mathbf{x}) - \mathbf{x}^{\ell,k}) \quad (\mathbf{x} \in \Omega). \quad (7)$$

The rate at which block (ℓ, k) approaches its target is $a_k M(\mathbf{x})$, a level-independent factor times the level exponent. Near a data point M grows like the inverse distance to it (Lemma B.4), and this growth is what makes the hitting time finite.

All blocks share one scalar speed M and relax towards a point of their own simplex at the level rate a_k . In the time ς with $d\varsigma = M(\mathbf{X}_s) ds$ the flow reads $d\mathbf{x}^{\ell,k}/d\varsigma = a_k(\bar{\mathbf{p}}^{\ell,k} - \mathbf{x}^{\ell,k})$, and the original time is finite iff $\int^\infty d\varsigma/M < \infty$. Near a vertex $\mathbf{e}_{j^*} \in \mathcal{M}_1^+$ we measure the distance by

$$V_{j^*}(\mathbf{x}) := \max_{\ell,k} \|\mathbf{x}^{\ell,k} - \mathbf{e}_{j_{\ell,k}^*}\|^{1/a_k}, \quad (8)$$

which weighs each level by its own schedule. Let $N := L \sum_k a_k v_k$ as in Lemma B.3.

The proof of convergence has a local and a global part. Locally, a trajectory that comes close to a data point must be absorbed in finite time, although $\dot{\alpha}_k(t) \rightarrow 0$ as $t \rightarrow 1$ when $a_k > 1$; globally, almost every trajectory from the prior must come close to some data point.

Theorem 3.7 (Local absorption). *Let $\mathbf{e}_{j^*} \in \mathcal{M}_1^+$. There are constants $\varepsilon \in (0, 1)$ and $0 < c \leq C$, depending on π_{j^*} , $(a_k, \sigma_k, v_k)_k$ and L , such that for every $\mathbf{x}_0$ with $0 < V_{j^*}(\mathbf{x}_0) < \varepsilon$:*

- (i) *the solution of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$, $\mathbf{X}_0 = \mathbf{x}_0$, stays in $\{V_{j^*} < \varepsilon\}$, $s \mapsto V_{j^*}(\mathbf{X}_s)$ is strictly decreasing, and $\mathbf{X}_s \rightarrow \mathbf{e}_{j^*}$ at a finite time $\tau(\mathbf{x}_0)$;*
- (ii) *if $N > 1$, then $V(\mathbf{x}_0) - Cs \leq V(\mathbf{X}_s) \leq V(\mathbf{x}_0) - cs$ for $s < \tau(\mathbf{x}_0)$, where $V = V_{j^*}$; hence*

$$\frac{V_{j^*}(\mathbf{x}_0)}{C} \leq \tau(\mathbf{x}_0) \leq \frac{V_{j^*}(\mathbf{x}_0)}{c}, \quad \|\mathbf{X}_s^{\ell,k} - \mathbf{e}_{j_{\ell,k}^*}\| \leq (C(\tau(\mathbf{x}_0) - s))^{a_k}. \quad (9)$$

The schedule exponent therefore does not slow absorption down: the time average in Eq. (1) produces the speed $M \asymp 1/V_{j^*}$, which blows up at the vertex. By Eq. (9), in the last stretch before absorption a level with a larger exponent is polynomially closer to its vertex than a level with a smaller one.

Theorem 3.7 settles the last stretch of a trajectory. For the global part the proof does not follow single trajectories; it works with the current $q\mathbf{b}$, whose divergence $\mu_0 - \mu_1$ (Section 2) already records that mass leaves the prior and arrives at the data (Appendix B.6).

Theorem 3.8 (Convergence and transport). Under [Assumption 3.2](#), with any $a_k > 0$, $\sigma_k > 0$ and sizes $v_k \geq 2$:

- (i) for μ_0 -almost every $\mathbf{x}$ the solution of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$, $\mathbf{X}_0 = \mathbf{x}$, reaches $\mathcal{M}_1^+$ at a finite time $\tau(\mathbf{x})$, and $\mathbb{E}_{\mathbf{x} \sim \mu_0}[\tau(\mathbf{x})] \leq 1$;
- (ii) the set $\mathcal{A} := \{\mathbf{x} \in \Omega : \tau(\mathbf{x}) < \infty\}$ is open and has full Lebesgue measure, and the map $\mathbf{T}(\mathbf{x}) := \lim_{s \uparrow \tau(\mathbf{x})} \mathbf{X}_s(\mathbf{x})$ is locally constant on $\mathcal{A}$; in particular $\mathbf{T} \in C^\infty(\mathcal{A})$ and $\mathbf{J}_{\mathbf{T}} \mathbf{b} = \mathbf{0}$ on $\mathcal{A}$;
- (iii) $\mathbf{T}_\# \mu_0 = \mu_1$;
- (iv) (occupation) for every Borel $\phi \geq 0$, $\mathbb{E}_{\mathbf{x} \sim \mu_0} \int_0^{\tau(\mathbf{x})} \phi(\mathbf{X}_s(\mathbf{x})) ds \leq \int_{\mathbb{R}^D} \phi q d\mathbf{x}$.

Numerically the bound $\mathbb{E}\tau \leq 1$ is attained: integrating the exact field from 1,000 prior samples gives $\mathbb{E}\tau = 1.000 \pm 0.014$ on Δ^2 and 1.005 ± 0.004 on $\Delta^2 \times \Delta^1$ ([Remark E.2](#)).

[Theorem 3.8](#) contains flat DBTM ($K = 1$) as a special case and holds for every exponent $a_k > 0$, so it covers both exponent and noise staggering. It does not claim absorption for *every* starting point (the field has stationary points, e.g. at symmetric configurations), and it does not address uniqueness of solutions of the conservation equation [Eq. \(2\)](#); it shows that the one-step map exists, transports μ_0 to μ_1 , and satisfies [Eq. \(2\)](#) on an open set of full measure.

For any block sizes and schedules, the map that DBTM learns is therefore well defined on almost all of the space, transports the prior to the data, and satisfies the conservation equation, and the mean hitting time is bounded by one. The bound on the occupation time in (iv) is used again in [Theorem 5.2](#).

4 When Does the Interpolant Commit?

We say that the interpolant has committed when the posterior probability of the true endpoint given $\mathbf{I}_t$ is close to one. The commitment time matters in two places: DBTM switches on its anchor loss after it, and its refinement sampler commits positions whose prediction is confident. We first describe the posterior, then locate the commitment time for a single simplex, then for the levels of a code, and finally show that the flow resolves the levels in the same order.

4.1 Factorisation of the posterior

Commitment is a property of the posterior of the endpoint given $\mathbf{I}_t$. For a code this posterior lives on $\mathcal{J} = \prod_{\ell,k} [v_k]$, and it can be studied one simplex at a time only if it factorises; the next proposition says when it does.

Proposition 4.1 (Factorisation). Fix $t \in (0, 1)$. If $\mu_1 = \bigotimes_{\ell,k} \mu_1^{\ell,k}$ factorises over positions and levels, then the posterior of the endpoint given $\mathbf{I}_t = \mathbf{x}$ factorises: $\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j \mid \mathbf{I}_t = \mathbf{x}, t) = \prod_{\ell,k} \mathbb{P}(\mathbf{x}_1^{\ell,k} = \mathbf{e}_{j_{\ell,k}} \mid \mathbf{I}_t^{\ell,k} = \mathbf{x}^{\ell,k}, t)$. If moreover $\mu_1^{\ell,k}$ is uniform on $[v_k]$, each factor is a softmax with inverse temperature $\beta_k(t) = (1 - \alpha_k(t))/(\alpha_k(t)^2 \sigma_k^2)$:

$$\mathbb{P}(\mathbf{x}_1^{\ell,k} = \mathbf{e}_i \mid \mathbf{I}_t^{\ell,k} = \mathbf{x}^{\ell,k}, t) = \frac{\exp(\beta_k(t) x_i^{\ell,k})}{\sum_{i'=1}^{v_k} \exp(\beta_k(t) x_{i'}^{\ell,k})}. \quad (10)$$

The factorisation holds at fixed t . The autonomous field mixes over t ([Proposition 3.3](#)) and does not factorise in general, because all levels share the same interpolation time.

4.2 A sharp commitment threshold

The anchor and the schedule rules below need the commitment time of blocks with a few dozen vertices, where a formula derived in the limit $V \rightarrow \infty$ gives no control of the error. We therefore prove two-sided

bounds that hold at every n , starting from an exact expression for the posterior mass of the truth.

Consider one block: a simplex with $n \geq 2$ vertices, uniform prior over vertices, true vertex $\mathbf{e}_{i^*}$, prior scale σ and schedule value $\alpha = \alpha(t) \in (0, 1)$. Writing $\mathbf{I}_t = (1 - \alpha)\mathbf{e}_{i^*} + \alpha\sigma\mathbf{g}$ with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, the posterior mass of the truth is a function of $\mathbf{g}$ and of the *signal-to-noise ratio* $\rho := (1 - \alpha)/(\alpha\sigma)$ only:

$$\pi_\star = \frac{1}{1 + \sum_{j \neq i^*} \exp(\rho(g_j - g_{i^*}) - \rho^2)} \quad (11)$$

([Lemma C.2](#)). Let $m = n - 1$ and $s_n = \sqrt{2 \log m}$.

In [Eq. \(11\)](#) the true vertex has an advantage of ρ^2 in log-odds over each competitor, and competitor j has the random score ρg_j . Commitment occurs when the advantage exceeds the largest of the $n - 1$ scores. Since the maximum of m independent standard Gaussians concentrates at $\sqrt{2 \log m}$, this happens when ρ crosses s_n . [Theorem 4.2](#) turns this heuristic into two bounds that hold for every n .

Theorem 4.2 (Sharp commitment threshold). (a) Commitment above the threshold. For all $u, u' > 0$ and $\rho \geq s_n + 2u$,

$$\mathbb{P}\left(1 - \pi_\star \leq \exp(u' - \rho(\rho - s_n - 2u))\right) \geq 1 - 2e^{-u^2/2} - e^{-u'}. \quad (12)$$

(b) No commitment below the threshold. For all $u > 0$, $h \in (0, s_n)$ and $\eta > 0$ with $\rho \leq s_n - h - u - \eta$,

$$\mathbb{P}(\pi_\star \leq e^{-\rho\eta}) \geq 1 - e^{-u^2/2} - e^{-\Lambda_n(h)}, \quad \Lambda_n(h) = \frac{s_n - h}{1 + (s_n - h)^2} \frac{e^{s_n h - h^2/2}}{\sqrt{2\pi}}. \quad (13)$$

Above the threshold the error $1 - \pi_\star$ decays like $e^{-\rho(\rho - s_n)}$; below it, $\pi_\star$ itself is exponentially small in ρ . The transition window has width of order one on the ρ scale while s_n grows like $\sqrt{\log n}$, so relative to its location the transition becomes sharp for large vocabularies. DBTM derives the location in the limit $V \rightarrow \infty$ from the random-energy model; the theorem gives it with explicit failure probabilities at finite n .

In the proof ([Appendix C](#)), part (a) bounds the sum in [Eq. \(11\)](#) from above by a truncated first moment, and part (b) bounds it from below by its largest term, which is where $\Lambda_n(h)$ comes from.

Corollary 4.3 (Asymptotic sharpness). Let $n \rightarrow \infty$ and fix $\varepsilon \in (0, 1)$, $c > 0$. If $\rho_n \geq (1 + \varepsilon)s_n$ then $\pi_\star \rightarrow 1$ in probability; if $c \leq \rho_n \leq (1 - \varepsilon)s_n$ then $\pi_\star \rightarrow 0$ in probability.

Corollary 4.4 (Commitment time). For $\alpha(t) = (1 - t)^a$ the signal-to-noise ratio $\rho(t) = ((1 - t)^{-a} - 1)/\sigma$ is a strictly increasing bijection $[0, 1) \rightarrow [0, \infty)$ with inverse $t(\rho) = 1 - (1 + \sigma\rho)^{-1/a}$. The threshold $\rho = s_n$ is crossed at

$$t^*(n, a, \sigma) = 1 - (1 + \sigma\sqrt{2 \log(n - 1)})^{-1/a}, \quad (14)$$

which is increasing in n and σ and decreasing in a . By [Theorem 4.2](#) the interpolant has committed with high probability for $t \geq t(s_n + 2u + \delta)$ and has not committed for $t \leq t(s_n - h - u - \eta)$.

[Equation \(14\)](#) is the DBTM formula [[47](#), Thm. 4.1] with $\log V$ replaced by $\log(n - 1)$, which agrees to leading order. [Theorem 4.2](#) adds the two-sided, finite- n guarantee and the width of the window: in the ρ scale the transition has width $O(u + h)$ with failure probabilities $e^{-u^2/2}$ and $e^{-\Lambda_n(h)}$, while s_n itself grows like $\sqrt{\log n}$, so the window becomes relatively sharp only slowly ([Fig. 3](#)).

4.3 Staggered commitment

A code is useful only if its levels become identifiable at different times. With product-uniform data, [Proposition 4.1](#) reduces every block to the single-simplex setting, so [Theorem 4.2](#) applies block by block with $n = v_k$

and $\rho = \rho_k(t)$, and what remains is to order the resulting thresholds.

Theorem 4.5 (Staggered commitment). *Assume $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$. For each block (ℓ, k) the posterior mass of its true vertex, $\pi_{\star}^{\ell,k}(t)$, has the law of Eq. (11) with $n = v_k$ and $\rho = \rho_k(t)$, where $\rho_k(t) = \frac{1 - \alpha_k(t)}{\alpha_k(t)\sigma_k}$, and the blocks are independent. Consequently Theorem 4.2 holds for every block with its own threshold*

$$t_k^{\star} = 1 - \left(1 + \sigma_k \sqrt{2 \log(v_k - 1)}\right)^{-1/a_k}, \quad (15)$$

and, by a union bound, all L positions of level k satisfy Eq. (12) simultaneously with probability at least $1 - L(2e^{-u^2/2} + e^{-u'})$. If $\sigma_k \sqrt{2 \log(v_k - 1)}$ is nondecreasing in k , $a_1 > a_2 > \dots > a_K$ and $v_k \geq 3$ for $k \geq 2$, then $t_1^{\star} < t_2^{\star} < \dots < t_K^{\star}$.

Each block competes against $v_k - 1$ rather than $V - 1$ alternatives, and on its own schedule. In the running example, the interpolant towards *network* commits to the cluster $\{\text{model}, \text{network}, \text{graph}\}$ at $t_1^{\star}$ and to *network* within it at $t_2^{\star} > t_1^{\star}$.

Corollary 4.6 (Staggering through the noise scale). *With linear schedules $a_k = 1$, $v_k \geq 3$, and scales such that $\sigma_k s_k$ is strictly increasing in k , where $s_k = \sqrt{2 \log(v_k - 1)}$, the levels commit at the strictly increasing times $t_k^{\star} = \sigma_k s_k / (1 + \sigma_k s_k)$.*

4.4 The flow inherits the level schedules

The results above concern the interpolant, which defines the field and the training distribution, whereas samples are produced by the map, that is, by the flow. Since the field averages over t , the ordering of the levels along the interpolant does not automatically carry over to the flow. This subsection shows that it does.

Theorem 4.5 concerns the posterior along the interpolant. We now show that the autonomous flow itself carries the level schedules, with no assumption on the data for the exact part. By Lemma 3.6 every block is pulled towards a point $\bar{\mathbf{p}}^{\ell,k}$ of its own simplex at the rate $a_k M(\mathbf{x})$, with one scalar M shared by all blocks. Along a solution $\mathbf{X}_s$ of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$ define the *flow time* $\varsigma(s) := \int_0^s M(\mathbf{X}_r) dr$ and $\mathbf{Z}(\varsigma) := \mathbf{X}_{s(\varsigma)}$.

Theorem 4.7 (Noise memory of the flow). *Under Assumption 3.2, let $\mathbf{x}_0 \notin \mathcal{M}_1$ and let $\mathbf{X}_s$, $s \in [0, S)$, be the maximal solution of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$ in $\mathbb{R}^D \setminus \mathcal{M}_1$. Then ς is a C^1 increasing bijection of $[0, S)$ onto $[0, \varsigma_{\max})$, and for every block and every $\varsigma \in [0, \varsigma_{\max})$*

$$\mathbf{Z}^{\ell,k}(\varsigma) = \lambda_k(\varsigma) \mathbf{x}_0^{\ell,k} + (1 - \lambda_k(\varsigma)) \mathbf{y}^{\ell,k}(\varsigma), \quad \lambda_k(\varsigma) = e^{-a_k \varsigma} = \alpha_k(1 - e^{-\varsigma}), \quad (16)$$

with $\mathbf{y}^{\ell,k}(\varsigma) \in \Delta^{v_k-1}$. If $\mathbf{1}^\top \mathbf{x}_0^{\ell,k} \neq 1$ for some block, which holds almost surely under μ_0 , then $\varsigma_{\max} = \infty$; the solution never reaches $\mathcal{M}_1$ in finite flow time, and $\text{dist}(\mathbf{Z}^{\ell,k}(\varsigma), \Delta^{v_k-1}) \leq \lambda_k(\varsigma)(\|\mathbf{x}_0^{\ell,k}\| + 1) \rightarrow 0$ for every block.

In the flow time ς , each block is a convex combination of its initial value and a point of its simplex, and the weight of the initial value is $e^{-a_k \varsigma}$, which is $\alpha_k(t)$ at $t = 1 - e^{-\varsigma}$. The flow thus removes the prior noise of each level at the rate of that level's schedule, for any data distribution.

At flow time ς the state is the interpolant formula with a point $\mathbf{y}$ of the polytope in place of the data and with exactly the prior noise $\alpha_k(t) \mathbf{x}_0^{\ell,k}$ that the interpolant carries at the common time $t = 1 - e^{-\varsigma}$. The flow therefore removes the prior noise of the levels on the schedules α_k themselves.

Corollary 4.8 (Level-wise denoising along the flow). *In the setting of Theorem 4.7: (i) $\lambda_k = \lambda_{k'}^{a_k/a_{k'}}$ for all levels; when a level with exponent a_c retains a fraction ε of its prior noise, a level with exponent $a_f < a_c$ retains the larger fraction ε^{a_f/a_c} . (ii) The ratio of the simplex part to the residual noise, $\bar{\rho}_k(\varsigma) := (1 - \lambda_k)/(\lambda_k \sigma_k)$, equals $\rho_k(1 - e^{-\varsigma})$; it crosses the threshold s_{v_k} of Theorem 4.2 at the flow time*

$$\varsigma_k^* = \frac{\ln(1 + \sigma_k \sqrt{2 \log(v_k - 1)})}{a_k} = -\ln(1 - t_k^*), \quad (17)$$

so the levels cross their thresholds along the flow in the same order as along the interpolant (Theorem 4.5).

Theorem 4.7 holds for any data distribution, but it does not say which vertex a block approaches; the point $\mathbf{y}^{\ell,k}$ is an average of the targets $\bar{\mathbf{p}}^{\ell,k}$. For symmetric data the decision is fixed at the start.

Theorem 4.9 (The flow keeps the order of the coordinates). *Fix a block (ℓ, k) and symbols $i \neq i'$, and assume that μ_1 is invariant under exchanging the symbols i and i' at block (ℓ, k) . Then the hyperplane $\{x_i^{\ell,k} = x_{i'}^{\ell,k}\}$ is invariant under the flow, and along every solution in $\mathbb{R}^D \setminus \mathcal{M}_1$ the sign of $X_{s,i}^{\ell,k} - X_{s,i'}^{\ell,k}$ is constant. If $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$, then for μ_0 -almost every $\mathbf{x}_0$: (i) the strict order of the coordinates of every block is preserved, so $i_{\ell,k}^* := \arg \max_i x_{0,i}^{\ell,k}$ is the largest coordinate of $\mathbf{X}_s^{\ell,k}$ for all s ; (ii) $\bar{\mathbf{p}}^{\ell,k}(\mathbf{X}_s)$ and $\mathbf{y}^{\ell,k}(\varsigma)$ have their largest coordinate at $i_{\ell,k}^*$; (iii) if $\mathbf{X}_s$ converges to a point of $\mathcal{M}_1$, the limit is $(\mathbf{e}_{i_{\ell,k}^*})_{\ell,k}$. In particular, wherever the transport map exists it is the blockwise argmax, $\mathbf{T}(\mathbf{x}_0)^{\ell,k} = \mathbf{e}_{\arg \max_i x_{0,i}^{\ell,k}}$, which pushes μ_0 to μ_1 .*

The proof uses only symmetry: exchanging i and i' maps the field to itself, so the hyperplane $\{x_i^{\ell,k} = x_{i'}^{\ell,k}\}$ is invariant and, by uniqueness of solutions, no trajectory crosses it. For product-uniform data this makes the one-step map explicit, and Theorem 6.8 uses this description to show what a narrow input layer cannot compute.

The order in time is controlled by the noise memory. Let $\mathbf{e}_{\ell,k}^* := \mathbf{e}_{i_{\ell,k}^*}$, let $\varsigma_{\ell,k}^\varepsilon := \inf\{\varsigma : \|\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}_{\ell,k}^*\| \leq \varepsilon\}$ be the flow time at which block (ℓ, k) arrives within ε of its vertex, and let

$$K_{\ell,k} := \int_0^\infty a_k e^{a_k \varsigma} \|\bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma)) - \mathbf{e}_{\ell,k}^*\| d\varsigma \in [0, \infty] \quad (18)$$

measure how fast the target of the block commits to that vertex, relative to the schedule of the level.

Theorem 4.10 (Arrival times). *Assume $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$ and let $\mathbf{x}_0$ satisfy the almost-sure conditions of Theorems 4.7 and 4.9. For every block and every $\varsigma \geq 0$,*

$$\lambda_k(\varsigma) \frac{|\mathbf{1}^\top \mathbf{x}_0^{\ell,k} - 1|}{\sqrt{v_k}} \leq \|\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}_{\ell,k}^*\| \leq \lambda_k(\varsigma) (\|\mathbf{x}_0^{\ell,k} - \mathbf{e}_{\ell,k}^*\| + K_{\ell,k}). \quad (19)$$

Consequently

$$\frac{1}{a_k} \ln \frac{|\mathbf{1}^\top \mathbf{x}_0^{\ell,k} - 1|}{\varepsilon \sqrt{v_k}} \leq \varsigma_{\ell,k}^\varepsilon \leq \max \left\{ 0, \frac{1}{a_k} \ln \frac{\|\mathbf{x}_0^{\ell,k} - \mathbf{e}_{\ell,k}^*\| + K_{\ell,k}}{\varepsilon} \right\}, \quad (20)$$

and if, for two levels c and f of the same position, the upper bound for c is smaller than the lower bound for f , then block (ℓ, c) arrives within ε of its vertex strictly before block (ℓ, f) , both in flow time and in the original time s .

The lower bound in Eq. (20) is unconditional. The upper bound is informative when $K_{\ell,k} < \infty$, i.e. when the target $\bar{\mathbf{p}}^{\ell,k}$ commits faster than the schedule $e^{-a_k \varsigma}$; then the arrival flow time of level k is $\frac{1}{a_k} \ln \frac{1}{\varepsilon} + O(1)$ as

$\varepsilon \rightarrow 0$, and the arrival flow times of exponent-staggered levels separate by the factor a_c/a_f . Noise staggering ($a_c = a_f$) separates them only by the additive amount $\frac{1}{a} \ln(\sigma_f/\sigma_c) + O(1)$.

Corollary 4.11 (Probability of coarse-before-fine arrival). *In the setting of Theorem 4.10, fix a position ℓ , levels $c \neq f$, $\varepsilon > 0$, $R > 0$, $r > \varepsilon\sqrt{v_f}$ and $\bar{K} \geq 0$. If*

$$\frac{a_c}{a_f} > \frac{\ln((R+1+\bar{K})/\varepsilon)}{\ln(r/(\varepsilon\sqrt{v_f}))}, \quad (21)$$

$$\text{then } \mathbb{P}(\varsigma_{\ell,c}^\varepsilon < \varsigma_{\ell,f}^\varepsilon) \geq 1 - \frac{2r}{\sqrt{2\pi v_f} \sigma_f} - \frac{v_c \sigma_c^2}{R^2} - \mathbb{P}(K_{\ell,c} > \bar{K}).$$

The local analysis of Section 3.3 shows that $K_{\ell,k} < \infty$ for product-uniform data.

Proposition 4.12 (The targets commit fast). *Assume $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$ with all $v_k \geq 2$. For μ_0 -almost every $\mathbf{x}_0$, $K_{\ell,k} < \infty$ for every block. More precisely, along the trajectory there are $\varsigma_1 \geq 0$ and $C < \infty$ with $V_{j^*}(\mathbf{Z}(\varsigma)) \leq Ce^{-\varsigma}$ and $\|\bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma)) - \mathbf{e}_{\ell,k}^*\| \leq Ce^{-N\varsigma}$ for $\varsigma \geq \varsigma_1$, where $\mathbf{e}_{j^*} = \mathbf{T}(\mathbf{x}_0)$ and $N = L \sum_k a_k v_k \geq 2a_{\max}$.*

Corollary 4.13 (The flow resolves coarse levels first). *Assume $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$ and $a_c > a_f$ for two levels c, f . For μ_0 -almost every $\mathbf{x}_0$ there is $\varepsilon_0(\mathbf{x}_0) > 0$ such that for every $\varepsilon \in (0, \varepsilon_0)$ and every position ℓ , block (ℓ, c) comes within ε of its vertex strictly before block (ℓ, f) , in flow time and in the original time s . Moreover, in Corollary 4.11 $\mathbb{P}(K_{\ell,c} > \bar{K}) \rightarrow 0$ as $\bar{K} \rightarrow \infty$.*

This is the flow-level counterpart of Theorem 4.5. It is visible in Fig. 4, where particles first collapse onto the cluster edges and then move along them to the words.

Remark 4.14 (Scope). Theorem 4.7 and Corollary 4.8 hold for any data. Theorem 4.9, Proposition 4.12 and Corollary 4.13 use the symmetry of product-uniform data; for correlated data the decision of a block is not fixed at time zero, and we do not have an ordering result. Noise staggering ($a_c = a_f$) separates the arrival times only by an additive $\frac{1}{a} \ln(\sigma_f/\sigma_c) + O(1)$, so Corollary 4.13 does not apply to it; numerically it orders the levels in 76% of trajectories (Appendix C.10).

4.5 Correlated data

The previous results assume that levels and positions are independent. In language the word depends on its cluster, some code combinations never occur, and neighbouring tokens are dependent. The threshold survives these correlations, with a shift that depends on how strongly the context can change the odds between two symbols.

Theorem 4.5 assumes that the data factorise over positions and levels. Real codes violate this twice: the fine symbol depends on the cluster (and some code configurations are unused), and positions are correlated. We now remove the assumption. Fix t and a block $B = (\ell, k)$ with $n = v_k$ vertices, $\alpha = \alpha_k(t)$, $\sigma = \sigma_k$, $\rho = \rho_k(t)$ and $\beta = \beta_k(t)$. Write a configuration as $\mathbf{j} = (i, o)$ with $i = j_{\ell,k}$ the symbol of B and $o = \mathbf{j}_{-B}$ the symbols of all other blocks (the *context*), $\pi(i, o) := \pi_{\mathbf{j}}$, $\pi_B(i) := \sum_o \pi(i, o)$, and for contexts o of positive mass $\pi(i | o) := \pi(i, o) / \sum_{i'} \pi(i', o)$ and $S(o) := \{i : \pi(i, o) > 0\}$. Let i^* be the true symbol and $\mathbf{g} := \mathbf{x}_0^{\ell,k} / \sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, so $\mathbf{I}_t^{\ell,k} = (1 - \alpha)\mathbf{e}_{i^*} + \alpha\sigma\mathbf{g}$.

A weighted threshold. Correlation enters the posterior of B only through positive weights on the competitors. We first extend Theorem 4.2 to weighted sums.

Theorem 4.15 (Weighted commitment threshold). Let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, $J \subseteq [n] \setminus \{i^*\}$ with $m := |J| \geq 1$, weights $w_j \in (0, \infty)$ for $j \in J$, and

$$\Sigma_w := \sum_{j \in J} w_j e^{\rho(g_j - g_{i^*}) - \rho^2}, \quad \pi_\star = \frac{1}{1 + \Sigma_w}. \quad (22)$$

Let $\bar{w} := \frac{1}{m} \sum_{j \in J} w_j$ and $s_{m+1} = \sqrt{2 \log m}$.

(a) For all $u, u' > 0$ and $\rho \geq s_{m+1} + 2u$: $\mathbb{P}(1 - \pi_\star \leq \Sigma_w \leq \bar{w} \exp(u' - \rho(\rho - s_{m+1} - 2u))) \geq 1 - 2e^{-u^2/2} - e^{-u'}$.

(b) For every $J' \subseteq J$ with $m' := |J'| \geq 2$ and $\hat{w} := \min_{j \in J'} w_j$, and all $u > 0$, $h \in (0, s_{m'+1})$, $\eta > 0$ with $\rho \leq s_{m'+1} - h - u - \eta$: $\mathbb{P}(\pi_\star \leq e^{-\rho\eta}/\hat{w}) \geq 1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$.

The same holds conditionally on any σ -algebra $\mathcal{F}$ independent of $\mathbf{g}$, with $\mathcal{F}$ -measurable weights.

For $w_j \equiv 1$ this is [Theorem 4.2](#). The weights shift the threshold only logarithmically: $\bar{w}e^{-\rho(\rho-s)} = 1$ at $\rho_w = \frac{1}{2}(s + \sqrt{s^2 + 4 \log \bar{w}}) = s + \frac{\log \bar{w}}{s} + O(s^{-3})$.

Side information. For general μ_1 the posterior of block B is exactly of this form.

Theorem 4.16 (Commitment with side information). For an arbitrary data distribution μ_1 , let $L(o) := \prod_{B' \neq B} \exp(\beta_{B'} x_{o_{B'}}^{B'})$, where $\beta_{B'}$ is the inverse temperature of block B' and $\mathbf{x} = \mathbf{I}_t$. For $j \in S_B := \{i : \pi_B(i) > 0\}$,

$$\log \frac{\mathbb{P}(j_{\ell,k} = j \mid \mathbf{I}_t)}{\mathbb{P}(j_{\ell,k} = i^* \mid \mathbf{I}_t)} = \rho(g_j - g_{i^*}) - \rho^2 + \Delta_j, \quad \Delta_j := \log \frac{\sum_o \pi(j, o) L(o)}{\sum_o \pi(i^*, o) L(o)}, \quad (23)$$

and $\mathbb{P}(j_{\ell,k} = j \mid \mathbf{I}_t) = 0$ for $j \notin S_B$. The offsets Δ_j are measurable with respect to $\mathcal{F}_{-B} := \sigma(\mathbf{x}_1, (\mathbf{x}_0^{B'})_{B' \neq B})$, which is independent of $\mathbf{g}$. Hence [Theorem 4.15](#) applies conditionally on $\mathcal{F}_{-B}$ with $J = S_B \setminus \{i^*\}$ and $w_j = e^{\Delta_j}$.

The dependence of block B on all other blocks and positions is thus summarised by one offset per competitor, which enters [Theorem 4.15](#) as a weight. The proof is Bayes' rule at fixed t : the likelihood of each block is proportional to $e^{\beta x_i}$, and summing over the contexts produces Δ_j .

Two ways of bounding the offsets cover the cases of interest. The first treats the context as a whole; the second marginalises a set F of strongly coupled blocks (for example the fine block under a coarse one) and pays for their evidence.

Proposition 4.17 (Bounded offsets). Let F be a set of blocks not containing B , write $o = (f, r)$ with f the symbols of the blocks in F , and $\pi(i, r) := \sum_f \pi(i, f, r)$. Assume rectangular support: $\pi(i, r) > 0$ for all $i \in S_B$ and all r with $\sum_i \pi(i, r) > 0$, and let $\Gamma_{B|F} := \max_r \max_{i, i' \in S_B} \pi(i, r) / \pi(i', r)$. With the evidence range $D_F := \sum_{B' \in F} (\max_i \beta_{B'} x_i^{B'} - \min_i \beta_{B'} x_i^{B'})$,

$$|\Delta_j| \leq \log \Gamma_{B|F} + D_F \quad \text{for all } j \in S_B.$$

For $F = \emptyset$, $D_\emptyset = 0$ and $\Gamma_B := \Gamma_{B|\emptyset} = \max_o \max_{i, i'} \pi(i \mid o) / \pi(i' \mid o)$. Moreover, for every block B' with v' vertices and every $u > 0$, $\mathbb{P}(\max_i \beta_{B'} x_i^{B'} - \min_i \beta_{B'} x_i^{B'} \leq \rho_{B'}^2 + 2\rho_{B'}(\sqrt{2 \log v'} + u)) \geq 1 - 2e^{-u^2/2}$.

When the support of B depends on the context (unused code configurations, a cluster that fixes the valid words), the offsets are unbounded and we condition on the context instead.

Proposition 4.18 (Mixture over contexts). Let $\mathcal{O}^* := \{o : \pi(i^*, o) > 0\}$ and, for $o \in \mathcal{O}^*$, $w_j(o) := \pi(j | o) / \pi(i^* | o)$ for $j \in S(o) \setminus \{i^*\}$. Then

$$\pi_{\star}^{\ell,k} = \sum_{o \in \mathcal{O}^*} \mathbb{P}(o | \mathbf{I}_t) \pi_{\star}(o), \quad \pi_{\star}(o) := \left(1 + \sum_{j \in S(o) \setminus \{i^*\}} w_j(o) e^{\rho(g_j - g_{i^*}) - \rho^2}\right)^{-1},$$

where $\mathbb{P}(o | \mathbf{I}_t)$ is the posterior of the context. Consequently, for every $\mathcal{G} \subseteq \mathcal{O}^*$, $1 - \pi_{\star}^{\ell,k} \leq \mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t) + \sup_{o \in \mathcal{G}} (1 - \pi_{\star}(o))$ and $\pi_{\star}^{\ell,k} \leq \sup_{o \in \mathcal{O}^*} \pi_{\star}(o)$.

Theorem 4.19 (Commitment under correlation). In the setting of [Proposition 4.18](#), let $\mathcal{G} \subseteq \mathcal{O}^*$, $S_{\mathcal{G}} := \bigcup_{o \in \mathcal{G}} S(o) \setminus \{i^*\}$, $m := |S_{\mathcal{G}}|$, $\Gamma_{\mathcal{G}} := \sup_{o \in \mathcal{G}} \max_j w_j(o)$, and let $J' := \bigcap_{o \in \mathcal{O}^*} S(o) \setminus \{i^*\}$, $m' := |J'|$, $\Gamma' := \sup_{o \in \mathcal{O}^*} \max_{j \in J'} w_j(o)^{-1}$. Conditionally on $\mathbf{x}_1$:

(a) if $m \geq 1$ and $\rho \geq s_{m+1} + 2u$, then with probability at least $1 - 2e^{-u^2/2} - e^{-u'}$,

$$1 - \pi_{\star}^{\ell,k} \leq \mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t) + \Gamma_{\mathcal{G}} \exp(u' - \rho(\rho - s_{m+1} - 2u));$$

(b) if $m' \geq 2$ and $\rho \leq s_{m'+1} - h - u - \eta$, then with probability at least $1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$, $\pi_{\star}^{\ell,k} \leq \Gamma' e^{-\rho\eta}$.

The context term $\mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t)$ is itself a posterior error of the other blocks, so the theorem can be applied recursively from coarse to fine. For two levels this gives staggered commitment without independence.

Corollary 4.20 (Staggered commitment under correlation). Let $K = 2$, let cluster c have the valid word indices $S(c) \subseteq [v_2]$, $n_c := |S(c)|$, and let the positions be arbitrarily dependent. Fix a position ℓ with true token (c^*, w^*) , write r for the symbols at all other positions, and assume

(H1) $\Gamma_1 := \max_r \max_{c, c'} \mathbb{P}(c_{\ell} = c | r) / \mathbb{P}(c_{\ell} = c' | r) < \infty$;

(H2) $\Gamma_2 := \max_{c, r} \max_{w, w' \in S(c)} \mathbb{P}(w_{\ell} = w | c_{\ell} = c, r) / \mathbb{P}(w_{\ell} = w' | c_{\ell} = c, r) < \infty$,

the maxima running over contexts of positive mass. Let $\rho_1 = \rho_1(t)$, $\rho_2 = \rho_2(t)$ and $\bar{D}_2 := \rho_2^2 + 2\rho_2(\sqrt{2\log v_2} + u)$. Then, with $\pi_{\star}^{\text{tok}} := \mathbb{P}(c_{\ell} = c^*, w_{\ell} = w^* | \mathbf{I}_t)$:

(i) cluster commits: if $\rho_1 \geq s_{v_1} + 2u$, then with probability at least $1 - 4e^{-u^2/2} - e^{-u'}$, $1 - \pi_{\star}^{\ell,1} \leq \Gamma_1 e^{\bar{D}_2} \exp(u' - \rho_1(\rho_1 - s_{v_1} - 2u))$;

(ii) token commits after its cluster: if $n_{c^*} \geq 2$ and $\rho_2 \geq s_{n_{c^*}} + 2u$, then with probability at least $1 - 2e^{-u^2/2} - e^{-u'}$, $1 - \pi_{\star}^{\text{tok}} \leq (1 - \pi_{\star}^{\ell,1}) + \Gamma_2 \exp(u' - \rho_2(\rho_2 - s_{n_{c^*}} - 2u))$;

(iii) the word is not yet decided: with $J' = \bigcap_{c: w^* \in S(c)} S(c) \setminus \{w^*\}$ and $m' = |J'| \geq 2$, if $\rho_2 \leq s_{m'+1} - h - u - \eta$, then with probability at least $1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$, $\pi_{\star}^{\ell,2} \leq \Gamma_2 e^{-\rho_2\eta}$.

In particular the cluster has committed and the word has not, simultaneously, at every t with $\rho_1(t)(\rho_1(t) - s_{v_1} - 2u) \geq \log \Gamma_1 + \bar{D}_2 + u' + \log(1/\delta)$ and $\rho_2(t) \leq s_{m'+1} - h - u - \eta$.

The offset caused by correlation enters the threshold condition as $\log \Gamma$, so the threshold moves by about $(\log \Gamma)/s$, which is small when s is large. Part (ii) is the situation created by refinement: once the cluster of a position has been fixed, the word competes only against the $n_c - 1$ valid words of that cluster.

Compared with [Theorem 4.5](#), correlation costs a factor Γ in the error, i.e. a shift of each threshold by $(\log \Gamma)/s$ on the ρ scale, and the fine level competes only against the $n_{c^*} - 1$ valid words of its cluster. Part (ii) is the statement used by refinement: once the cluster is committed, $\mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t)$ vanishes.

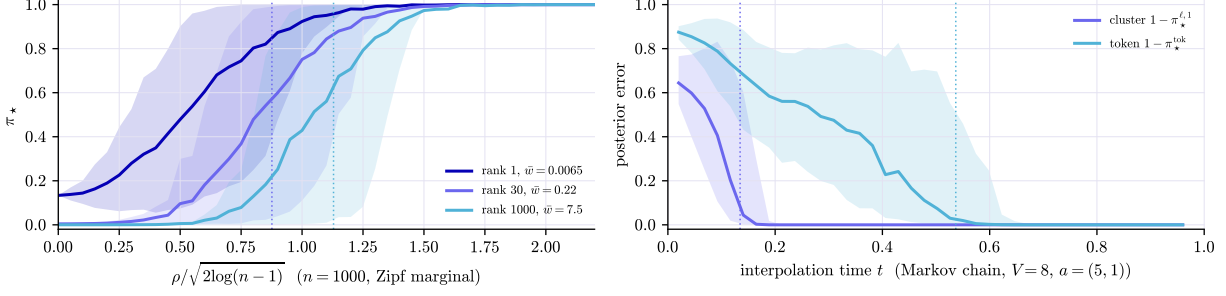


Figure 2 *Commitment under correlation.* Left: posterior mass of the truth for a Zipf marginal over $n = 1000$ symbols (mean and 10–90% band, 400 draws), for a frequent, a middle and a rare true symbol; dotted lines are the crossovers ρ_w of Theorem 4.15. Right: exact posterior error of the cluster and of the token at the middle position of a three-position Markov chain over $V = 8$ tokens (clusters of sizes 3, 3, 2, $a = (5, 1)$, $\sigma = 0.55$), median and 10–90% band over 400 draws; dotted lines are the worst-case crossovers $\rho(\rho - s) = \log \Gamma$ of Corollary 4.20 with the exact Γ_1, Γ_2 and $s = s_{v_1}, s_{v_2}$.

Corollary 4.21 (Anchoring under correlation). Under (H1)–(H2), Proposition 5.5(ii) holds with the error bounds of Corollary 4.20: for the cluster block, $\mathbb{E}[1 - \pi_*^{\ell,1}] \leq \Gamma_1 e^{\bar{D}_2} e^{u' - \rho_1(\rho_1 - s_{v_1} - 2u)} + 4e^{-u^2/2} + e^{-u'}$, and for the token, conditionally on the true token (c^*, w^*) , $\mathbb{E}[1 - \pi_*^{\text{tok}} \mid \mathbf{x}_1] \leq \mathbb{E}[1 - \pi_*^{\ell,1} \mid \mathbf{x}_1] + \Gamma_2 e^{u' - \rho_2(\rho_2 - s_{n_{c^*}} - 2u)} + 2e^{-u^2/2} + e^{-u'}$.

Remark 4.22 (Non-uniform symbols). For product data with non-uniform block marginals p , Eq. (23) gives $\Delta_j = \log(p_j/p_{i^*})$ exactly, so $\bar{w} = (1 - p_{i^*})/(m p_{i^*})$ with $m = |S_B| - 1$. Frequent symbols ($\bar{w} < 1$) commit before s and rare ones after it, at $\rho_w \approx s + \log \bar{w}/s$; for a Zipf marginal over 1000 symbols the predicted crossovers $\rho_w/s = 0.88$ (rank 30) and 1.13 (rank 1000) match the simulated ones, 0.82 and 1.04 (Fig. 2, left).

Remark 4.23 (Markov sources). If the tokens form a Markov chain with all transition probabilities in $[a, b]$, $a > 0$, then at an interior position $\Gamma_1 \leq (n_{\max}/n_{\min})(b/a)^2$ and $\Gamma_2 \leq (b/a)^2$, where $n_{\max}, n_{\min}$ are the largest and smallest cluster sizes. For a toy chain with $V = 8$ tokens in clusters of sizes 3, 3, 2 (Fig. 2, right) the exact constants are $\Gamma_1 = 4.3$ and $\Gamma_2 = 7.1$; across 16,000 draws the sandwich of Proposition 4.17 and the inequality of Proposition 4.18 hold in every draw, and the bounds of Corollary 4.20(i)–(ii) with $u = 3$, $u' = 4$, evaluated with the realised evidence range D_2 and without the cluster term, are never violated (0/11,600 and 0/2,918 draws in the regime where they apply).

4.6 Limits of the hierarchy

The staggered thresholds compare the levels of one code with each other. They do not say whether the code changes the time at which the whole token is determined, and a fair comparison with flat DBTM must separate the effect of the hierarchy from the extra signal energy that the stacked vertices carry.

A coarse-to-fine token is a stack of K one-hot vectors and has squared norm K , against 1 for a flat token, so at equal σ it carries K times the signal energy. The next result compares the two at equal energy.

Proposition 4.24 (Energy-matched comparison). Scale every level vertex by $c > 0$, so a token has squared norm $c^2 K$. Then the signal-to-noise ratio of block (ℓ, k) is $c \rho_k(t)$. Take equal sizes $v_k = v \geq 2$, $V = v^K$, equal schedules and scales, and match the energy of a flat token ($c = K^{-1/2}$). The finest level then crosses its threshold at $\rho = \sqrt{2K \log(v-1)}$ and the flat token at $\rho = \sqrt{2 \log(V-1)}$, and

$$\sqrt{2K \log(v-1)} \leq \sqrt{2 \log(V-1)}, \quad \lim_{v \rightarrow \infty} \frac{K \log(v-1)}{\log(v^K - 1)} = 1. \quad (24)$$

So at matched energy and schedules the hierarchical token is fully resolved no later than a flat token, and at the same time to leading order. What changes is *when partial information becomes available*: coarse levels commit before the finest one, which is what refinement and anchoring can use.

5 Training and Sampling

We now turn to a learned map $\mathbf{T}_\theta$. The questions are whether the DBTM loss still has the right minimiser in the new coordinates, how its value relates to the quality of one-step samples, and how anchoring, refinement and the choice of schedules should account for levels that commit at different times.

Transport loss. The DBTM transport loss is applied to the stacked state,

$$\mathcal{L}_{\text{transport}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \sum_{\ell, k} \left\| \mathbf{T}_\theta^{\ell, k}(\mathbf{I}_t) - \text{sg}(\mathbf{T}_\theta^{\ell, k}(\mathbf{I}_t) + \mathbf{J}_{\mathbf{T}_\theta^{\ell, k}}(\mathbf{I}_t) \dot{\mathbf{I}}_t) \right\|^2, \quad (25)$$

where $\mathbf{J}_{\mathbf{T}_\theta^{\ell, k}}(\mathbf{x}) \in \mathbb{R}^{v_k \times D}$ is the Jacobian of block (ℓ, k) of the map. The sum over blocks equals the squared norm of the stacked residual, so this is exactly the DBTM loss in the new coordinates.

Proposition 5.1 (Regression step). *Fix a map $\bar{\mathbf{T}} \in C^1$ with $\mathbb{E} \|\bar{\mathbf{T}}(\mathbf{I}_t) + \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2 < \infty$. Over measurable maps $\mathbf{T}$ with $\mathbb{E} \|\mathbf{T}(\mathbf{I}_t)\|^2 < \infty$, the minimiser of $\mathbb{E} \|\mathbf{T}(\mathbf{I}_t) - \bar{\mathbf{T}}(\mathbf{I}_t) - \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2$ is*

$$\mathbf{T}^*(\mathbf{x}) = \bar{\mathbf{T}}(\mathbf{x}) + \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{x}) \mathbf{b}(\mathbf{x}) \quad \text{for almost every } \mathbf{x}, \quad (26)$$

blockwise $\mathbf{T}^{*, \ell, k} = \bar{\mathbf{T}}^{\ell, k} + \mathbf{J}_{\bar{\mathbf{T}}^{\ell, k}} \mathbf{b}$. A continuous $\bar{\mathbf{T}}$ is a fixed point, $\mathbf{T}^* = \bar{\mathbf{T}}$, if and only if $\mathbf{J}_{\bar{\mathbf{T}}} \mathbf{b} = \mathbf{0}$ on $\mathbb{R}^D \setminus \mathcal{M}_1$, i.e. if and only if it satisfies the conservation equation Eq. (2) off the data.

Each step regresses the network onto a copy of itself advanced by one Euler step along the field. The minimiser is the old map advanced along $\mathbf{b}$, and the map is a fixed point exactly when it satisfies the conservation equation. The loss is a sum over blocks and needs no modification for the coarse-to-fine coordinates.

From training loss to sample quality. DBTM characterises the minimiser of its transport loss but gives no bound that links a nonzero loss to the error of the learned map [47]. We prove such a bound. It controls the distance between the one-step sample $\mathbf{T}_\theta(\mathbf{x}_0)$ and the exact sample $\mathbf{T}(\mathbf{x}_0)$ by two quantities: the boundary error on the data and the residual of the conservation equation Eq. (2) under the time-marginal law of the interpolant, which is the law on which the transport loss is evaluated.

Throughout, $\Omega = \mathbb{R}^D \setminus \mathcal{M}_1^+$ (Section 3.3), q is the density of $\mathbf{I}_t$ with $t \sim \mathcal{U}[0, 1]$ (Lemma B.3), and for a measurable $\mathbf{f}$ we write $\mathbb{E}_q \mathbf{f} := \int \mathbf{f}(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \mathbf{f}(\mathbf{I}_t)$.

By Theorem 3.8, for μ_0 -almost every $\mathbf{x}$ the solution $\mathbf{X}_s(\mathbf{x})$ of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$, $\mathbf{X}_0 = \mathbf{x}$, stays in Ω on $[0, \tau(\mathbf{x}))$ with $\tau(\mathbf{x}) < \infty$, the limit $\mathbf{T}(\mathbf{x}) \in \mathcal{M}_1^+$ exists, and $\mathbf{T}_{\#} \mu_0 = \mu_1$. We use these facts throughout.

The *occupation measure* of the flow started from the prior is

$$\nu(A) := \mathbb{E}_{\mathbf{x} \sim \mu_0} \int_0^{\tau(\mathbf{x})} \mathbb{1}_{\{\mathbf{X}_s(\mathbf{x}) \in A\}} ds, \quad A \subseteq \Omega \text{ Borel}, \quad (27)$$

so that $\int \mathbf{f} d\nu = \mathbb{E}_{\mu_0} \int_0^\tau \mathbf{f}(\mathbf{X}_s) ds$ for every measurable $\mathbf{f} \geq 0$ (Tonelli), and $\nu(\Omega) = \mathbb{E}_{\mu_0} \tau$.

Theorem 5.2 (Training loss controls sample error). *Let $\mathbf{T}_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^D$ be continuous on $\mathbb{R}^D$ and C^1 on Ω , and define the boundary error and the conservation residual*

$$\varepsilon_{\text{bnd}}(\theta) := \mathbb{E}_{\mathbf{y} \sim \mu_1} \|\mathbf{T}_\theta(\mathbf{y}) - \mathbf{y}\|, \quad R(\theta) := \mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\| = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \|\mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \mathbf{b}(\mathbf{I}_t)\|.$$

$$(i) \quad \mathbb{E}_{\mathbf{x} \sim \mu_0} \|\mathbf{T}_\theta(\mathbf{x}) - \mathbf{T}(\mathbf{x})\| \leq \varepsilon_{\text{bnd}}(\theta) + \int \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\| d\nu.$$

(ii) Consequently,

$$W_1((\mathbf{T}_\theta)_\# \mu_0, \mu_1) \leq \mathbb{E}_{\mu_0} \|\mathbf{T}_\theta - \mathbf{T}\| \leq \varepsilon_{\text{bnd}}(\theta) + R(\theta). \quad (28)$$

(iii) The residual is the excess of the DBTM transport loss over its minimum in one regression step: with $\mathcal{E}(\mathbf{T}; \mathbf{T}) := \mathbb{E} \|\mathbf{T}(\mathbf{I}_t) - \mathbf{T}(\mathbf{I}_t) - \mathbf{J}_{\mathbf{T}}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2$ and $\mathbb{E} \|\mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2 < \infty$,

$$R(\theta)^2 \leq \mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\|^2 = \mathcal{E}(\mathbf{T}_\theta; \mathbf{T}_\theta) - \min_{\mathbf{T}} \mathcal{E}(\mathbf{T}; \mathbf{T}_\theta) \leq \mathbb{E} \|\mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2, \quad (29)$$

where the minimum is over measurable $\mathbf{T}$ with $\mathbb{E} \|\mathbf{T}(\mathbf{I}_t) - \mathbf{T}_\theta(\mathbf{I}_t)\|^2 < \infty$, and the right-hand side is the value of the transport loss Eq. (25) at θ .

(iv) If every block of $\mathbf{T}_\theta$ is $c \cdot \text{softmax}$ of logits (vertices scaled by $c > 0$), then $\varepsilon_{\text{bnd}}(\theta) \leq 2c \mathbb{E}_{\mathbf{x}_1} \sum_{\ell, k} \text{CE}(\mathbf{T}_\theta^{\ell, k}(\mathbf{x}_1)/c, \mathbf{x}_1^{\ell, k}/c)$, i.e. $2c$ times the boundary loss summed over blocks.

The proof follows a trajectory of the exact flow: along it, $\mathbf{T}_\theta$ changes by the residual $\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}$, so the distance between the learned and the exact end point is at most the integral of the residual plus the boundary error at the end. The quantity that enters is the residual, that is, how much one further regression step would lower the loss, and not the loss value, which also contains a variance term that remains positive at the exact map (Remark 5.4).

Corollary 5.3 (Decoding error). Under the hypotheses of Theorem 5.2(ii) and with vertices scaled by $c > 0$, decode a state by the per-block argmax, $\text{dec}(\mathbf{x})^{\ell, k} := \arg \max_i x_i^{\ell, k}$ (ties broken arbitrarily), and then by γ^{-1} . Then

- (i) for every block, $\mathbb{P}(\text{dec}(\mathbf{T}_\theta(\mathbf{x}_0))^{\ell, k} \neq \text{dec}(\mathbf{T}(\mathbf{x}_0))^{\ell, k}) \leq \frac{\sqrt{2}}{c} \mathbb{E} \|\mathbf{T}_\theta^{\ell, k}(\mathbf{x}_0) - \mathbf{T}^{\ell, k}(\mathbf{x}_0)\|$;
- (ii) the expected number of wrongly decoded tokens is at most $\frac{\sqrt{2K}}{c} \sum_{\ell} \mathbb{E} \|\mathbf{T}_\theta^{\ell}(\mathbf{x}_0) - \mathbf{T}^{\ell}(\mathbf{x}_0)\| \leq \frac{\sqrt{2KL}}{c} (\varepsilon_{\text{bnd}} + R)$, where $\mathbf{T}^{\ell}$ stacks the K blocks of position ℓ ;
- (iii) the law of the decoded one-step sample is within total variation $\min\{1, \frac{\sqrt{2KL}}{c} (\varepsilon_{\text{bnd}} + R)\}$ of the data distribution.

Remark 5.4 (What the bound does and does not say). (a) The transport loss value $\mathbb{E} \|\mathbf{J}_{\mathbf{T}_\theta} \dot{\mathbf{I}}_t\|^2$ does not vanish at the exact map: it equals the residual $\mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\|^2$ plus the variance $\mathbb{E} \|\mathbf{J}_{\mathbf{T}_\theta}(\dot{\mathbf{I}}_t - \mathbf{b})\|^2$, which is positive wherever the map is not locally constant. The quantity that the bound needs is the excess Eq. (29), i.e. how much one more regression step could still lower the loss; the loss value is only an upper bound. (b) The bound is a statement about one-step sampling; refinement (Algorithm 1) changes the input distribution and is not covered. (c) The margin factor $\sqrt{2K}/c$ in Corollary 5.3 equals $\sqrt{2}$ for flat DBTM and $\sqrt{2}K$ for an energy-matched code ($c = K^{-1/2}$), but the residuals of the two models are measured in different coordinates, so the corollary is not a comparison between them. (d) The bound uses only Theorem 3.8; no assumption beyond those of Assumption 3.2 is needed.

Level-wise anchor. DBTM anchors $\mathbf{T}_\theta(\mathbf{I}_t)$ to $\mathbf{x}_1$ by cross-entropy after a single commitment time. With a hierarchy, each level has its own threshold, so we anchor level k after t_k^* :

$$\mathcal{L}_{\text{anchor}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \sum_{\ell, k} \mathbb{1}[t > t_k^*] \text{CE}(\mathbf{T}_\theta^{\ell, k}(\mathbf{I}_t), \mathbf{x}_1^{\ell, k}), \quad (30)$$

where $\mathbf{T}_\theta^{\ell, k}$ is parameterised as a softmax over v_k classes. The full objective is $\mathcal{L} = \mathcal{L}_{\text{transport}} + \lambda_{\text{bnd}} \mathcal{L}_{\text{bnd}} + \lambda_{\text{semi}} \mathcal{L}_{\text{semi}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}}$, with the boundary and semigroup losses of DBTM summed over blocks.

Proposition 5.5 (Anchoring is Bayes-consistent after the level's threshold). Fix t and a block (ℓ, k) , and let $\mathbf{p}^{\ell, k}(\mathbf{x})$ be the posterior of $\mathbf{x}_1^{\ell, k}$ given $\mathbf{I}_t = \mathbf{x}$.

- (i) If every symbol of block (ℓ, k) has positive probability under μ_1 , then among all maps $\mathbf{q}$ into the open simplex, $\mathbb{E} \text{CE}(\mathbf{q}(\mathbf{I}_t), \mathbf{x}_1^{\ell, k})$ is minimised exactly by $\mathbf{q} = \mathbf{p}^{\ell, k}$ almost everywhere.
- (ii) The anchor target differs from this minimiser by $\mathbb{E} \|\mathbf{p}^{\ell, k}(\mathbf{I}_t) - \mathbf{x}_1^{\ell, k}\|_1 = 2 \mathbb{E}[1 - \pi_\star^{\ell, k}]$. Under the assumptions of [Theorem 4.5](#), for $\rho_k(t) \geq s_{v_k} + 2u$, $\mathbb{E}[1 - \pi_\star^{\ell, k}] \leq e^{u' - \rho_k(\rho_k - s_{v_k} - 2u)} + 2e^{-u^2/2} + e^{-u'}$, while for $\rho_k(t) \leq s_{v_k} - h - u - \eta$, $\mathbb{E}[1 - \pi_\star^{\ell, k}] \geq (1 - e^{-\rho_k \eta})(1 - e^{-u^2/2} - e^{-\Lambda_{v_k}(h)})$.

The anchor target is the data vertex itself, which is the Bayes-optimal prediction only after the posterior has committed; before that, the optimal prediction is a distribution over several symbols. With a code the levels commit at different times, so the anchor is switched on separately for each level.

Before $t_k^\star$ the anchor would pull the predictor away from a non-degenerate posterior, which is the mode collapse DBTM observes when anchoring too early [47, Sec. 4.3]. A single anchor time is therefore too late for coarse levels or too early for fine ones. Under correlated data the same statements hold with the shifted thresholds of [Section 4.5](#) ([Corollary 4.21](#)).

Decision budget. The transport loss is an expectation over $t \sim \mathcal{U}[0, 1]$, so the share of training states in which a level is still undecided is the share of interpolation time before its threshold. Only those states teach the map the level's decision from noise.

Proposition 5.6 (Decision budget). Assume $\mu_1 = \bigotimes_{\ell, k} \text{Unif}([v_k])$ and write $t_k(\rho) := 1 - (1 + s_k \rho)^{-1/a_k}$, so that $t_k^\star = t_k(s_{v_k})$. For a block (ℓ, k) of a training pair $(t, \mathbf{I}_t)$, $t \sim \mathcal{U}[0, 1]$, and all $u, u', h, \eta, \delta > 0$ admissible in [Theorem 4.2](#) with $s_{v_k} - h - u - \eta \geq 0$,

$$\begin{aligned} \mathbb{P}(\pi_\star^{\ell, k} \leq e^{-\rho_k \eta}) &\geq t_k(s_{v_k} - h - u - \eta) (1 - e^{-u^2/2} - e^{-\Lambda_{v_k}(h)}), \\ \mathbb{P}(1 - \pi_\star^{\ell, k} \leq e^{u' - \rho_k \delta}) &\geq (1 - t_k(s_{v_k} + 2u + \delta)) (1 - 2e^{-u^2/2} - e^{-u'}). \end{aligned}$$

As $v_k \rightarrow \infty$ with fixed relative margins, the failure probabilities vanish and both time boundaries differ from $t_k^\star$ by $o(1)$, so the undecided share is $t_k^\star + o(1)$ and the decided share $1 - t_k^\star + o(1)$. Consequently exponent staggering with a fast coarse level ($a_1 > 1$) shrinks the budget of the coarse level, $t_1^\star = 1 - (1 + \sigma_1 s_{v_1})^{-1/a_1}$, which is decreasing in a_1 , whereas staggering by slowing the fine levels ($a_K < 1$) keeps $t_1^\star$ and enlarges $t_K^\star$.

For one-step sampling the map has to determine every level from pure noise, and for the coarse level this is the hardest step. Under uniform training times, a fraction of about $t_1^\star$ of the training pairs show the coarse level undecided, and exponent staggering with $a_1 > 1$ reduces this fraction. Staggering by slowing the fine levels keeps the order of commitment and leaves the coarse budget unchanged.

Coarse-to-fine refinement. DBTM refines a one-step sample by committing confident positions and renoising the rest [47, Sec. 5]. We commit at the level of blocks and respect the hierarchy: a block can commit only once the coarser blocks of its position have committed ([Algorithm 1](#)).

Proposition 5.7 (Refinement terminates and respects the hierarchy). In every round, either $|\Delta| \geq n_r$ or every block is committed; after round k all blocks are committed; and no block is ever committed before a coarser block of the same position. [Algorithm 1](#) uses exactly k evaluations of $\mathbf{T}_\theta$.

Because commitment is decided block by block, the sampler can fix the cluster of a position while its word is still uncertain; the next evaluation then chooses among the words of that cluster.

Committing whole tokens. [Algorithm 1](#) counts blocks. A sequence has LK of them, and a fine block can be committed only together with or after the coarser blocks of its position, so under a budget of k rounds the floor rule spends the early rounds on coarse blocks and leaves most fine blocks to the last rounds, where they are decided together and without each other as context. A variant keeps the unit of DBTM: each round commits whole positions, ranked by the joint confidence $q^\ell = \prod_k q_k^{\ell, k}$, with the threshold rule applied to q^ℓ

Algorithm 1 Coarse-to-fine refinement with k evaluations of T_θ

```
1:  $\mathbf{x} \leftarrow \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \Sigma)$ ;  $\mathcal{C}_0 \leftarrow \emptyset$  ▷ committed blocks
2: for  $r = 1, \dots, k$  do
3:    $\hat{\mathbf{x}} \leftarrow T_\theta(\mathbf{x})$ ;  $q^{\ell,k} \leftarrow \max_i \hat{x}_i^{\ell,k}$  for all blocks ▷ one NFE
4:    $\mathcal{R}_r \leftarrow$  uncommitted blocks;  $n_r \leftarrow \lceil |\mathcal{R}_r| / (k - r + 1) \rceil$ ;  $\Delta \leftarrow \emptyset$ 
5:   for  $k' = 1, \dots, K$  do ▷ threshold rule, coarse levels first
6:     add to  $\Delta$  every  $(\ell, k') \in \mathcal{R}_r$  with  $q^{\ell,k'} \geq \kappa$  whose coarser blocks lie in  $\mathcal{C}_{r-1} \cup \Delta$ 
7:   end for
8:   while  $|\Delta| < n_r$  do ▷ floor rule
9:     add to  $\Delta$  the frontier block with the largest  $q$ ; the frontier is, for each position, the coarsest block
       not in  $\mathcal{C}_{r-1} \cup \Delta$ 
10:  end while
11:   $\mathcal{C}_r \leftarrow \mathcal{C}_{r-1} \cup \Delta$ ;  $\mathbf{x}^{\ell,k} \leftarrow e_{\arg \max_i \hat{x}_i^{\ell,k}}$  for  $(\ell, k) \in \mathcal{C}_r$ ;  $\mathbf{x}^{\ell,k} \leftarrow$  fresh  $\mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I})$  draw otherwise
12: end for
13: return the per-block argmax of  $\hat{\mathbf{x}}$ , decoded by  $\gamma^{-1}$ 
```

and the floor rule $n_r = \lceil |\mathcal{R}_r| / (k - r + 1) \rceil$ counted in positions. Every level of a position is committed at once, so no word precedes its cluster, and the argument of [Proposition 5.7](#) shows that the variant terminates after exactly k evaluations. For a flat code the two rules coincide. In the trained models of [Table 3](#) the token-level rule is the better sampler for coarse-to-fine codes.

Refinement in the loop. At inference the context of round r is the set the sampler has committed, not a random set, and a map trained only on random contexts sees it out of distribution. DBTM therefore trains on the states the sampler visits [[47](#), Sec. 5.2]: it rolls out the sampler without gradients, pins the committed positions to the data, renoises the others, and adds the cross-entropy of T_θ against $\mathbf{x}_1$ on the uncommitted positions of every intermediate state. We use this objective with the token-level commit rule.

6 Commitment Profile, Schedule Rule and the Embedding Bottleneck

[Proposition 4.24](#) shows that the code adds no signal, and [Theorem 5.2](#) bounds the error of any learned map without saying which map is easier to learn. This section isolates three properties that a learner does see: the shape of a single commitment, the share of training time that a schedule spends on it, and how much of a token’s state a network can read through its input layer. The first two are statements about schedules and hold for flat DBTM as well; the third separates the flat and the coarse-to-fine state.

6.1 Profile of the commitment transition

[Theorem 4.2](#) locates the transition at $\rho = s_n$ with a window of width $O(1)$. The next result identifies the transition itself.

Theorem 6.1 (Commitment profile). *In the single-block setting of [Theorem 4.2](#), let $n \geq 3$, $s = s_n$, $g_\star := g_{i^\star}$, and write $\rho = s + x$ with $\rho > 0$. Let $\Lambda := \log(\pi_\star / (1 - \pi_\star))$ be the log-odds of the true vertex. For all $y > 0$ and $w \in (0, s)$, with probability at least $1 - e^{-y} - e^{-\Lambda_n(w)}$,*

$$\rho g_\star + s x + \frac{x^2}{2} - y \leq \Lambda \leq \rho g_\star + \rho(x + w), \quad (31)$$

and for every $\eta > 0$, with $\epsilon_n := e^{-y} + e^{-\Lambda_n(w)}$,

$$\Phi\left(x - \frac{x^2 + 2y}{2\rho} - \eta\right)(1 - e^{-\rho\eta}) - \epsilon_n \leq \mathbb{E} \pi_\star \leq \Phi(x + w + \eta) + e^{-\rho\eta} + \epsilon_n. \quad (32)$$

Consequently, as $n \rightarrow \infty$, $\sup_{|x| \leq X} |\mathbb{E} \pi_\star(s_n + x) - \Phi(x)| \rightarrow 0$ for every $X > 0$, and $\Lambda/s_n - (x + g_\star) \rightarrow 0$ in probability.

The proof writes the log-odds as $\Lambda = \rho g_\star + \rho^2 - \log S$, with S the sum over the $n - 1$ competitors, and bounds $\log S$ from above by Markov's inequality and from below by its largest term; what remains random is the Gaussian $g_\star$ of the true coordinate, and it produces Φ .

In the scale of the signal-to-noise ratio every commitment has the same profile, a standard normal distribution function centred at s_n : the mean posterior mass of the truth rises from 10% to 90% while ρ crosses $[s_n - 1.28, s_n + 1.28]$, for every vocabulary size. A larger vocabulary moves the profile and does not reshape it. The explicit bounds of Eq. (32) converge slowly, at rate $O(\log s_n/s_n)$, and are loose at realistic n ; in Monte Carlo the profile is already within 0.07 of $\Phi(x)$ at $n = 10^3$, shifted slightly to the left. A single sample commits much faster, over an interval of ρ of length $O(1/s_n)$ located at $s_n - g_\star$; the width of the population profile comes from the noise on the true coordinate, not from the competitors. We call $\{t : |\rho_k(t) - s_{v_k}| \leq z\}$ the *commitment window* of level k and its length $W_k(z) = t_k(s_{v_k} + z) - t_k(s_{v_k} - z)$ the level's *window share*: the fraction of training states, under $t \sim \mathcal{U}[0, 1]$, in which the posterior of the level is neither close to the prior nor close to a point mass. Throughout, $z = 1.28$ (the central 80% of the profile) and $s_{v_k} > z$.

6.2 Allocation of training time

Under $t \sim \mathcal{U}[0, 1]$ a level learns its decision from noise only on training states inside its commitment window, so the window share is the quantity a schedule should be designed for. A change of the time coordinate makes it explicit for every power schedule.

Proposition 6.2 (Gumbel coordinates). *Let $\ell(t) = \log(-\log(1 - t))$. If $t \sim \mathcal{U}[0, 1]$, then ℓ has the density $\varphi_G(\ell) = \exp(\ell - e^\ell)$ and the distribution function $G(\ell) = 1 - \exp(-e^\ell)$. For a schedule $\alpha(t) = (1 - t)^a$ with prior scale σ and $s > z > 0$, the window $\{t : |\rho(t) - s| \leq z\}$ is the interval $\ell \in [c, c + \lambda]$ with*

$$c = \log \frac{\log(1 + \sigma(s - z))}{a}, \quad \lambda = \lambda_s(\sigma) := \log \frac{\log(1 + \sigma(s + z))}{\log(1 + \sigma(s - z))}.$$

The length λ_s is a strictly decreasing bijection of $(0, \infty)$ onto $(0, L_s)$, $L_s := \log \frac{s+z}{s-z}$. Hence the windows of power schedules are exactly the intervals $[c, c + \lambda]$ with $c \in \mathbb{R}$ and $0 < \lambda < L_s$: the exponent moves the window, and the prior scale sets its length.

Theorem 6.3 (The largest window share). *For every power schedule, the window share of a level with threshold $s > z$ satisfies*

$$W(z) < M(L_s) := \sup_{c \in \mathbb{R}} (G(c + L_s) - G(c)) = \exp\left(-\frac{L_s}{e^{L_s} - 1}\right) - \exp\left(-\frac{L_s e^{L_s}}{e^{L_s} - 1}\right), \quad (33)$$

and the bound is approached as $\sigma \rightarrow 0$ with $a = (e^{L_s} - 1) \log(1 + \sigma(s - z))/L_s$; the optimal window contains the mode of φ_G , i.e. the time $t = 1 - 1/e$. Moreover $M(L) \leq L/e$, and since $L_s = 2z/s + O(s^{-3})$, a level with n vertices spends at most a fraction $\approx 2z/(e s_n)$ of the training time in its window. The linear schedule $a = 1$ reaches at most $W_{\text{lin}} = 2z/(\sqrt{s + z} + \sqrt{s - z})^2$, attained at $\sigma = (s^2 - z^2)^{-1/2}$, which is $\approx e/4 \approx 68\%$ of $M(L_s)$ for large s .

The proof is a one-dimensional maximisation: the share is $G(c + \lambda) - G(c)$, which increases in $\lambda < L_s$ and, as a function of the position c , is largest where the density φ_G takes the same value at both ends of the window. The bound is used twice below, for a single flat level (Corollary 6.4) and for several levels that share the time axis (Proposition 6.5).

For flat DBTM the prior scale is fixed in practice ($\sigma = 1$ in [47]), and only the exponent is free. Then the optimum is explicit.

Table 1 Window shares under the best z -ordered exponents ([Proposition 6.5](#)), prior scale $\sigma = 0.55$ for every level, $z = 1.28$. Flat: one level of size V with its best exponent. K levels of equal size $v \approx V^{1/K}$; every level receives the share shown. Exponents in parentheses, coarse first.

V	flat	$K = 2$	$K = 3$
256	0.188 ($a = 0.99$)	0.286 (1.16, 0.48)	0.222 (2.65, 0.66, 0.17)
1,024	0.161 (1.07)	0.247 (1.20, 0.58)	0.229 (2.14, 0.75, 0.27)
4,096	0.142 (1.14)	0.219 (1.23, 0.65)	0.221 (1.96, 0.81, 0.34)
16,384	0.128 (1.20)	0.198 (1.26, 0.72)	0.211 (1.86, 0.86, 0.40)
50,257	0.119 (1.24)	0.184 (1.29, 0.76)	0.202 (1.81, 0.89, 0.44)

Corollary 6.4 (A schedule rule for DBTM). Fix $n \geq 3$ vertices, a prior scale $\sigma > 0$ and $z \in (0, s_n)$, and let $\lambda = \lambda_{s_n}(\sigma)$. Over $a > 0$, the window share of the schedule $(1 - t)^a$ is maximised exactly at

$$a^*(n, \sigma) = \log(1 + \sigma(s_n - z)) \frac{e^\lambda - 1}{\lambda}, \quad (34)$$

where it equals $M(\lambda)$, and the commitment time $t^* = 1 - (1 + \sigma s_n)^{-1/a^*}$ lies in the optimal window, which contains $t = 1 - 1/e$. For $\sigma = 1$, $z = 1.28$: $a^* = 1.40, 1.50, 1.58, 1.70$ at $V = 256, 1,024, 4,096, 50,257$, with $t^* \approx 0.64$ in each case, whereas the linear schedule $a = 1$ commits at $t^* = 0.77, 0.79, 0.80, 0.82$ and has a window share smaller by 6%, 9%, 12%, 15%.

The rule says that DBTM’s decision should sit near $t = 1 - 1/e$ whatever the vocabulary; the linear schedule pushes it later as V grows. The anchor of [Eq. \(30\)](#) moves with it, to $t^*(a^*)$. [Section 7](#) tests the rule in training.

With several levels the windows compete for the same time axis. Call schedules z -ordered if the windows are disjoint and coarse first, $t_k(s_{v_k} + z) \leq t_{k+1}(s_{v_{k+1}} - z)$.

Proposition 6.5 (Staggering as interval packing). Fix prior scales σ_k and let $\lambda_k = \lambda_{s_{v_k}}(\sigma_k)$. The largest value m^* of $\min_k W_k(z)$ over z -ordered choices of the exponents a_k is the largest m for which the following greedy chain exists: $e_0 = -\infty$ and, for $k = 1, \dots, K$, $c_k = \max\{e_{k-1}, \inf C_k(m)\} \leq \sup C_k(m)$ and $e_k = c_k + \lambda_k$, where $C_k(m) = \{c : G(c + \lambda_k) - G(c) \geq m\}$ is an interval. The optimal exponents are $a_k = \log(1 + \sigma_k(s_{v_k} - z)) e^{-c_k}$, and $m^* \leq \min\{\min_k M(\lambda_k), 1/K\}$.

[Table 1](#) turns this into a design rule. The best codes place adjacent windows around $t = 1 - 1/e$, coarse first, with exponents close to one: a coarse exponent somewhat above one and fine exponents below one. The exponents of our figures, $a = (5, 1)$, put the coarse window below $t \approx 0.2$, where it receives a small share; this is the effect described by [Proposition 5.6](#). Slowing the fine levels too far pushes their windows into the thin right tail of φ_G and starves them in turn. Within the power family, every level of a two- or three-level code receives a larger share than a flat token of the same vocabulary. Staggering through the noise scale alone ([Corollary 4.6](#)) is the case $a_k = 1$, in which only the lengths of the windows differ.

Remark 6.6 (What the window share does not show). The window share is a property of the power family. A general increasing $\rho(t)$ can spend any fraction of the training time inside any window, flat DBTM included, so the share alone is not an advantage of the hierarchy. It is a design rule for the schedules that DBTM and this paper use, and it explains why a fast coarse schedule hurts.

6.3 The embedding bottleneck

A transformer reads the state of a position through a linear first layer, the soft analogue of a token embedding. DBTM’s backbone is a DiT of hidden size 768 acting on the (B, L, V) simplex state with untied input and output embeddings, at vocabularies $V = 30,522, 50,257$ and 49,152 [[47](#), App. E–G], so each token’s state passes through a linear map of rank at most $768 \ll V$. The coarse-to-fine state has $\sum_k v_k$ coordinates per

position, which fits through such a layer. We show that this difference is decisive for the map that DBTM is trained to learn. Consider a block with n vertices whose state $\mathbf{x} \in \mathbb{R}^n$ is read only through $W\mathbf{x}$, $W \in \mathbb{R}^{d \times n}$ of rank r ; whatever follows is arbitrary.

Theorem 6.7 (The bottleneck delays commitment). *In the single-block setting with a uniform prior over $n \geq 2$ vertices, every estimator $\hat{i} = \psi(W\mathbf{I}_t)$ with ψ measurable satisfies*

$$\mathbb{P}(\hat{i} \neq i^*) \geq 1 - \frac{\frac{r}{2} \log(1 + \rho^2/n) + \log 2}{\log n}. \quad (35)$$

Consequently an error at most ε requires $\rho^2 \geq \frac{2n}{r}((1 - \varepsilon) \log n - \log 2)$. Without a bottleneck ($r = n$) the bound requires only $\rho^2 \geq 2((1 - \varepsilon) \log n - \log 2)$, in agreement with s_n .

The proof is Fano's inequality: whatever follows the layer, the information that $W\mathbf{I}_t$ carries about i^* is at most that of an r -dimensional Gaussian channel, $\frac{r}{2} \log(1 + \rho^2/n)$.

A rank- r reader therefore cannot commit before $\rho \approx s_n \sqrt{n/r}$, a threshold polynomial in n instead of logarithmic. At DBTM's OpenWebText setting ($V = 50,257$, width 768) and with $\varepsilon = 0.1$, it needs $\rho \geq 34.4$ against $s_V = 4.65$, i.e. $t \geq 0.97$ instead of $t^* = 0.82$ for the linear schedule with $\sigma = 1$. Over the whole interval $t \in [t^*, 0.97]$ the DBTM anchor asks the network for a decision that no function of its input layer can make.

Theorem 6.7 limits decisions at a fixed time. The DBTM objective asks for more, namely the transport map, which for symmetric data is the argmax of the prior draw (**Theorem 4.9**); the next theorem bounds how often a rank- r reader can compute it.

Theorem 6.8 (The bottleneck hides the transport map). *Let $K = 1$ and μ_1 uniform on n vertices, so that $\mathbf{T}(\mathbf{x}_0) = \mathbf{e}_{\arg \max_i x_{0,i}}$ by **Theorem 4.9**. Let $\hat{i}(\mathbf{x}_0)$ be the vertex nearest to $f(W\mathbf{x}_0)$ for a measurable $f : \mathbb{R}^d \rightarrow \mathbb{R}^n$ and W of rank r . Then*

$$\mathbb{P}(\hat{i}(\mathbf{x}_0) = \arg \max_i x_{0,i}) \leq p_n(r) := \inf_{\theta \in (0,1], 0 \leq c \leq \tau} \left[\frac{r}{\theta n} + Q(\tau - c) + n Q(c/\sqrt{\theta}) + e^{-nQ(\tau)} \right], \quad (36)$$

and $\mathbb{E}\|f(W\mathbf{x}_0) - \mathbf{T}(\mathbf{x}_0)\| \geq (1 - p_n(r))/\sqrt{2}$. As $n \rightarrow \infty$, $p_n(r) \leq 2r \log n/n + o(1)$ uniformly in r .

The proof uses that the projection onto the row space of W has trace r , so at most r/θ coordinates keep a weight of at least θ , while the argmax of the prior draw is uniform over all n coordinates.

Corollary 6.9 (A floor for the flat objective). *In the setting of **Theorem 6.8**, every map $\mathbf{T}_\theta = f \circ W$ that satisfies the hypotheses of **Theorem 5.2** has $\varepsilon_{\text{bnd}}(\theta) + R(\theta) \geq (1 - p_n(r))/\sqrt{2}$. For a coarse-to-fine state read injectively (W injective on $\mathbb{R}^{\sum_k v_k}$), the transport map $\mathbf{T}(\mathbf{x}_0)^{\ell,k} = \mathbf{e}_{\arg \max_i x_{0,i}^{\ell,k}}$ is a function of $W\mathbf{x}_0$, and the obstruction vanishes.*

At $n = 50,257$ and $r = 768$, the setting of DBTM on OpenWebText, $p_n(r) \leq 0.157$: any network with a width-768 linear input layer gets the transport map of flat DBTM wrong on at least 84% of noise draws, and its boundary error plus conservation residual stays above 0.59 however long it is trained. Simulations put the best achievable value near $r/n \approx 1.5\%$, attained by a layer that reads 768 coordinates exactly; a random layer does worse (**Appendix F**). A two-level code with $v = 225$ per level has $450 \leq 768$ coordinates per position and is read without loss.

The theorem does not say that flat DBTM produces bad samples: a bottlenecked network can still push the prior to the data with a different map. It says that the map the DBTM objective defines is out of reach, so the objective has a floor and **Theorem 5.2** cannot certify the flat model, while for the coarse-to-fine state the obstruction is absent. The statement concerns the symmetric data model of **Theorem 4.9**; for structured

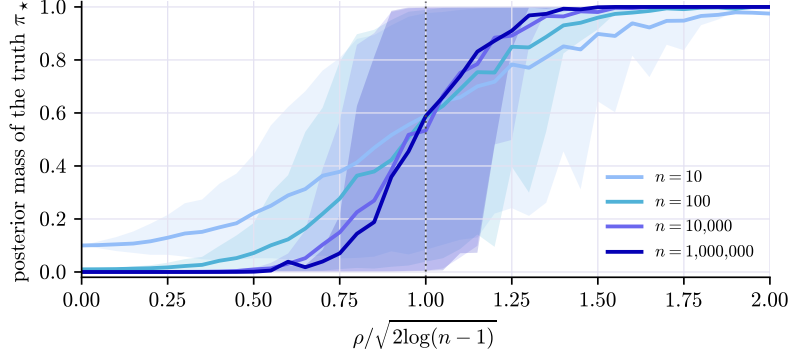


Figure 3 The commitment threshold sharpens with n . Monte Carlo estimate of $\mathbb{E}[\pi_\star]$ (lines) and its 10–90% range (bands) against $\rho/\sqrt{2\log(n-1)}$, from Eq. (11) with 400 draws per point (120 for $n = 10^6$). The transition is centred at the threshold of Theorem 4.2 and steepens slowly, as the $\sqrt{\log n}$ scaling predicts.

data the transport map is not the argmax, and only Theorem 6.7 applies directly. The bound of Theorem 6.8 becomes informative once $2r \log V/V$ is well below one, for instance $V > 10r \log V$; this is the regime in which the theorem predicts a handicap for flat DBTM and none for the coarse-to-fine state.

7 Experiments

The statements of Sections 3 to 6 concern exact quantities, so Section 7.1 checks them with exact posteriors and exact fields, without a network. Whether they matter for a trained map is a separate question: Section 7.2 trains flat DBTM and coarse-to-fine models on a synthetic language with known likelihood, and Section 7.3 measures the cost of a training step. Details are in Appendix G.

7.1 Exact computations

The threshold theorem. Figure 3 evaluates Eq. (11). At $\rho/s_n = 0.8, 1.0$ and 1.2 the mean posterior mass of the truth is 0.41, 0.59, 0.72 for $n = 10$ and 0.15, 0.59, 0.87 for $n = 10^6$: the crossover sits at the predicted threshold for every n and sharpens as n grows.

Staggered commitment. The bottom row of Fig. 1 reports the Monte Carlo posterior entropy of each level of a three-level code ($a = 3, 1.6, 1$, $\sigma = 1$) for $V = v^3$ from 64 to 32,768. The simulated transitions of each level fall on the times $t_k^\star$ of Eq. (15), in the predicted order. The flat curve is not energy-matched (see the caption of Fig. 1). Since the levels are independent under product data, the posterior entropy of a token is the sum of the entropies of its levels (Remark C.1), so the level curves add up to the token’s.

Exact fields. Figure 1 (top) shows the exact field of Proposition 3.3 on $\Delta^2 \times \Delta^2$ with $a_{\text{cluster}} = 5$, $a_{\text{word}} = 1$. Every streamline enters the word triangle of a cluster before it resolves a word, and the field is tangent to the affine hull, as Proposition 3.4 requires. Figure 4 integrates the flow from 900 prior samples: clusters resolve before words, and hitting times are finite on a slice through the polytope. Figure 5 shows the corresponding flat DBTM field.

Refinement. Figure 6 runs Algorithm 1 with the exact per-level posterior in place of $\mathbf{T}_\theta$ on the sentence “the map sends each point to the word” over a 3×3 vocabulary. Clusters commit in rounds 1–2 and words in rounds 2–4, and blocks below κ are committed by the floor rule.

7.2 Trained models

Setup. The language is a first-order Markov chain over $V = C \cdot W = 1,024$ tokens with $C = W = 32$, written as (cluster, word): the transition factorises as $P(c' | c)P(u' | c', u)$, and every cluster and every pair (cluster, previous word) has $S = 2$ successors, so each token has 4 valid successors among 1,024. Sequences

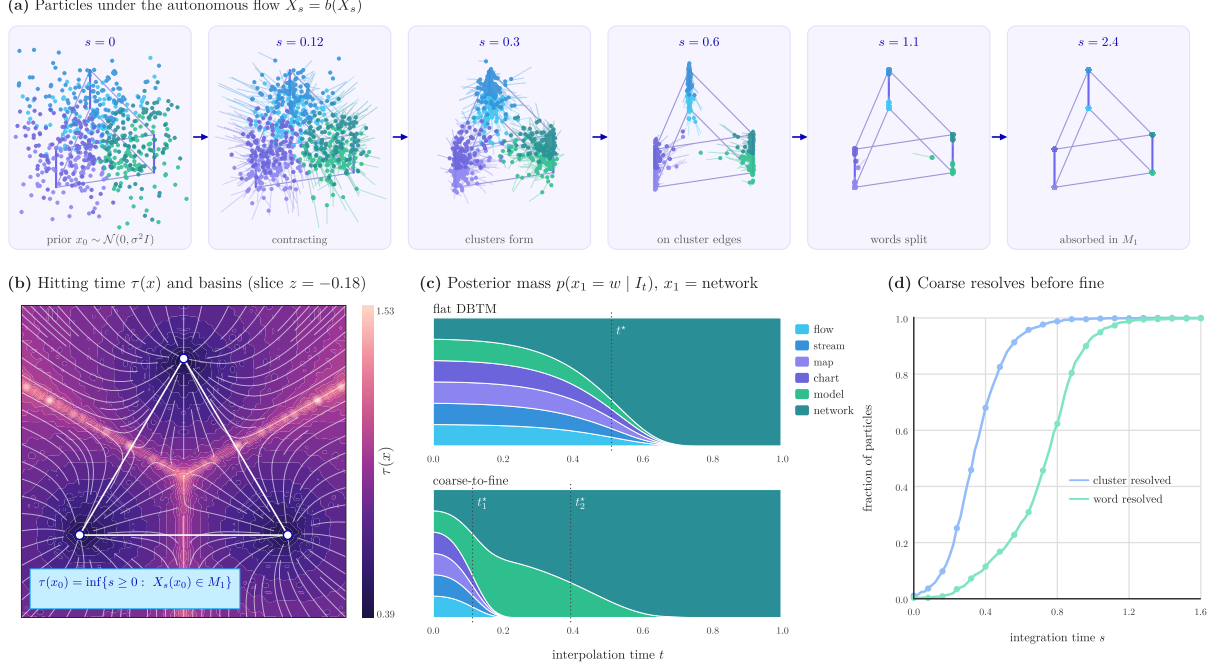


Figure 4 *Dynamics of the coarse-to-fine flow on $\Delta^2 \times \Delta^1$ ($a_{\text{cluster}} = 5$, $a_{\text{word}} = 1$, $\sigma = 0.55$). (a) 900 prior samples under $\dot{X}_s = b(X_s)$ collapse onto the cluster edges, then split into words. (b) Hitting time $\tau(x)$ on a slice. (c) Posterior mass $\mathbb{P}(x_1 = w | I_t)$ along the interpolant towards network: the flat token stays six-way uncertain longer (not energy-matched), the coarse-to-fine token first concentrates on its cluster. (d) Fraction of particles whose cluster and word are resolved.*

have length $L = 8$, and the exact likelihood gives the validity of every sampled transition and the negative log-likelihood (NLL) of every sample; the entropy rate is 1.89 nats per token. Every model is the same small transformer (3 layers, 4 heads) with a linear input layer on the state of each position and one softmax output per block. The flat model reads 1,024 coordinates per position at width 64; the coarse-to-fine model reads $32 + 32$ coordinates and, to match the parameter count (238k against 233k for flat DBTM), uses width 96. Both are trained for 30k steps with the transport, boundary and level-wise anchor losses (Eqs. (25) and (30)), $\sigma = 1$, energy-matched vertices for the code (Proposition 4.24), and are evaluated on 2,000 samples with $\kappa = 0.9$.

The schedule rule improves DBTM. Corollary 6.4 gives $a^* = 1.50$ for $V = 1,024$ and $\sigma = 1$, against the linear schedule of DBTM. With standard training the rule raises the valid transitions after four evaluations from 31.2% to 39.1% and lowers the NLL in every seed (Table 2). With refinement-in-loop training it raises the valid sequences after eight evaluations from 51.1% to 72.2%, i.e. it removes 43% of the invalid sequences. The rule changes one scalar of DBTM and nothing else: the network, the losses and the sampler are the same, and the anchor moves with t^* .

Coarse-to-fine against flat DBTM. At matched parameters the one-step samples of the two models are equally valid (Table 3, $k = 1$). The coarse level is resolved much better: with standard training ($a = (1, 1)$ against flat $a = 1$), 71% of the cluster transitions of a one-step coarse-to-fine sample are valid, against 29% for flat DBTM, and when the true tokens are given at every other position the coarse-to-fine map fills the gaps with 21.5% exact tokens against 15.1%. The model loses under refinement. With Algorithm 1 its samples barely improve with more evaluations (18.0% against 31.4% at $k = 4$): the fine blocks are left to the last rounds and decided without each other as context, as described in Section 5. Refinement-in-loop training together with the token-level rule brings it level with flat DBTM under the linear schedule at four evaluations, but flat DBTM with a^* stays ahead, gains more from further rounds, and at eight evaluations is clearly better. In this experiment the hierarchy therefore helps the coarse decision and the use of context;

Δ^2 : one token, $V = 3$

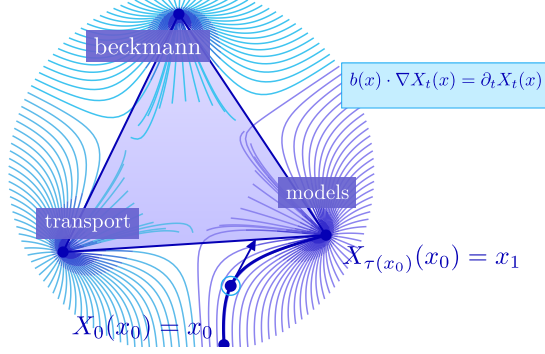


Figure 5 *Flat DBTM on one simplex Δ^2 ($V = 3$), computed exactly from Eq. (1) with the linear interpolant and $\sigma = 0.6$. Every streamline flows directly to one vertex. The highlighted trajectory $X_s(x_0)$ is an integrated streamline ending at models.*

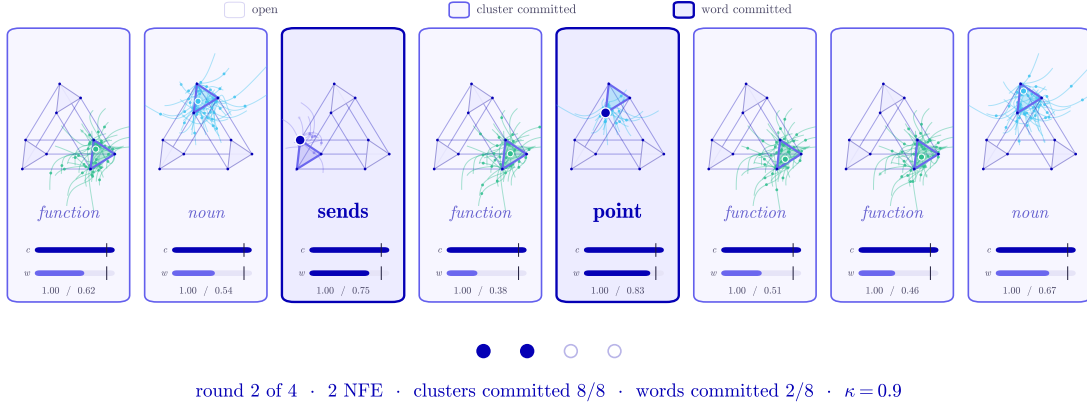


Figure 6 *Coarse-to-fine refinement (Algorithm 1) with exact per-level posteriors, after round 2 of 4. Each card is one token on its own duoprism $\Delta^2 \times \Delta^2$; the exact field is drawn inside and narrows to the committed cluster (violet word triangle) and then to the committed word. Bars give the cluster (c) and word (w) confidence with a tick at $\kappa = 0.9$.*

we conjecture that the remaining gap lies in how the fine level ranks its own proposals during refinement. The embedding bottleneck of Theorem 6.8 concerns the exact transport map; at $V = 1,024$ and $r = 64$, $2r \log V/V \approx 0.87$ and the bound is not yet informative, so the experiment does not test the regime in which the theorem has force. The flat model learns a different map that still generates valid sequences. The exponents $a = (1.92, 1.09)$ are the z -ordered packing of Proposition 6.5 for $v = 32$ at the effective prior scale $\sigma/c = \sqrt{2}$ of the energy-matched code; in training they do not improve on $a = (1, 1)$, so for the code the window rule is not confirmed by this experiment.

7.3 Cost of a training step

The coarse-to-fine state is small: at the GPT-2 vocabulary a position has $D = 111$ coordinates ($K = 3$, $v_k = 37$) instead of 50,257. The transport loss differentiates the network forward along $\dot{I}_t$ through the whole state, and the input and output layers scale with D , so the state size enters every training step. Table 4 measures one full training step with the same transformer trunk for every code. With a width-768, 12-layer trunk at $V = 50,257$, the coarse-to-fine step ($K = 3$) takes 326 ms against 804 ms (2.5 $\times$), uses 1.7 GB against 5.3 GB, has 57M against 134M parameters, and samples in 22 ms against 60 ms. With a small trunk, where the vocabulary layers dominate, the step is 14 $\times$ faster and the memory 18 $\times$ smaller. For the coarse-to-fine codes the cost is essentially independent of V ; for flat DBTM it grows with V . These are costs, not sample quality; at matched parameters the trained comparison is Table 3.

Table 2 *The schedule rule on flat DBTM* ($V = 1,024$, width 64, 233k parameters, $\sigma = 1$, 30k steps). Left: without refinement-in-loop training, mean $\pm$ standard deviation over 3 seeds; samples from Algorithm 1 after 1 and 4 evaluations. Right: with refinement-in-loop training (seed 0), token-level sampler after 4 and 8 evaluations. Valid transitions and sequences in %, NLL in nats per token under the true chain (data: 1.89).

schedule	standard training			refinement in the loop			
	trans. $k=1$	trans. $k=4$	NLL $k=4$	trans. $k=4$	trans. $k=8$	seq. $k=8$	NLL $k=8$
linear, $a = 1$ [47]	16.5 ± 0.8	31.2 ± 1.7	9.47 ± 0.19	56.5	88.2	51.1	2.92
Corollary 6.4, $a^* = 1.50$	20.6 ± 1.1	39.1 ± 0.7	8.58 ± 0.08	59.5	94.3	72.2	2.25

Table 3 *Coarse-to-fine against flat DBTM at matched parameters* ($V = 1,024 = 32 \times 32$, seed 0, 30k steps; flat width 64, coarse-to-fine width 96, energy-matched). Valid transitions (trans.) and sequences (seq.) in %. Standard training: Algorithm 1. Refinement in the loop: Algorithm 1 at $k = 4$, and the token-level rule at $k = 4, 8$.

model	params	standard training		refinement in the loop			
		trans. $k=1$	trans. $k=4$	blocks $k=4$	tokens $k=4$	tokens $k=8$	seq. $k=8$
flat DBTM, $a = 1$	233k	16.0	31.4	57.3	56.5	88.2	51.1
flat DBTM, $a^* = 1.50$	233k	20.8	39.5	59.7	59.5	94.3	72.2
coarse-to-fine, $a = (1, 1)$	238k	15.8	18.0	20.4	55.5	73.5	17.6
coarse-to-fine, $a = (1.92, 1.09)$	238k	15.2	16.5	19.5	53.9	74.6	18.6

8 Related Work

Discrete and simplex-based generative models. Discrete diffusion models corrupt tokens by masking or uniform noise [5, 29, 39, 42], and discrete flow matching learns continuous-time Markov chains [13, 18]. Continuous alternatives move probability vectors on or around the simplex [14, 16, 46], building on flow matching and stochastic interpolants [2, 3, 27, 28].

Few-step maps. Consistency and shortcut models [17, 45] and flow maps [8, 9] learn to jump along the probability-flow ODE; flow maps have been adapted to discrete sequences [24, 35, 38] and to refinement [1]. Beckmann transport models replace the time-dependent flow by an autonomous one and learn its endpoint map through a conservation equation [25]; DBTM instantiates them on the simplex for language [47]. Our work changes the state space and the schedules of DBTM and keeps its training principle.

Hierarchical token models. Class-based language models and hierarchical softmax factor a word into a class and a word within it [11, 19, 30, 31]. Residual quantisation and semantic identifiers code items as tuples of codewords [26, 36], and next-scale prediction generates images coarse to fine [23, 48]. Closest to us, HDLM [52] builds a hierarchical vocabulary for masked discrete diffusion in which tokens move to their ancestors in the forward process; follow-ups predict over a vocabulary tree [50] or diffuse tokens and clusters jointly in a continuous model [33]. Our construction differs in the generative mechanism: a time-independent field on a product of simplices with a one-step map, per-level schedules rather than transitions between vocabularies, and a commitment theory for the levels. A clustering such as HDLM’s could supply our code γ .

Phase transitions in generative models. Diffusion models commit to coarse features in sharp transitions [4, 6, 37, 51], and on hierarchical data the transition reveals the levels of the hierarchy [40]. DBTM locates the commitment time with the random-energy model [15]. Theorem 4.2 proves a two-sided, finite- n version of this threshold with elementary Gaussian bounds [10, 20], Theorem 4.5 orders the thresholds of a code, and Theorem 6.1 and Corollary 6.4 turn them into a schedule rule for DBTM.

9 Limitations and Conclusion

Limitations. (i) The trained experiments use one synthetic language with 1,024 tokens, small networks and up to three seeds; they show that the schedule rule improves DBTM in that setting, not on natural language at scale, where DBTM’s own evaluation uses vocabularies of $3\text{--}5 \cdot 10^4$ tokens and generative perplexity. (ii)

Table 4 *Cost of one training step* (transport loss with a forward-mode JVP, boundary and anchor losses, backward, AdamW; sequence length 128; Apple M5 GPU, one process per configuration). Same transformer trunk for every code; the codes differ only in the input and output layers. D : coordinates per position; step: median over 8 steps; memory: peak allocation; sample: one evaluation of the map.

V	code	D	params	step (ms)	memory (GB)	sample (ms)
<i>trunk: width 256, 4 layers, batch 8</i>						
4,096	flat	4,096	4.2M	67	1.20	4.6
4,096	$K = 2$	128	2.2M	37	0.25	2.5
4,096	$K = 3$	48	2.2M	36	0.25	2.6
16,384	flat	16,384	10.5M	164	2.23	11.4
16,384	$K = 2$	256	2.3M	37	0.26	2.7
16,384	$K = 3$	78	2.2M	36	0.25	2.5
50,257	flat	50,257	27.9M	504	4.46	36.0
50,257	$K = 2$	450	2.4M	39	0.29	2.7
50,257	$K = 3$	111	2.2M	37	0.25	2.5
<i>trunk: width 768, 12 layers, batch 4</i>						
4,096	flat	4,096	63.1M	347	2.15	23.8
4,096	$K = 2$	128	57.0M	318	1.72	21.7
4,096	$K = 3$	48	56.9M	316	1.71	21.5
16,384	flat	16,384	82.0M	449	3.60	31.0
16,384	$K = 2$	256	57.2M	319	1.73	21.6
16,384	$K = 3$	78	56.9M	316	1.72	21.7
50,257	flat	50,257	134.1M	804	5.30	59.8
50,257	$K = 2$	450	57.5M	316	1.76	22.1
50,257	$K = 3$	111	57.0M	326	1.72	22.0

The coarse-to-fine code of the experiments matches the clusters of the language exactly; codes built from data, e.g. by clustering embeddings, are not tested. (iii) The embedding bottleneck (Theorem 6.8) concerns the exact transport map of symmetric data and carries a $\log V$ factor that the simulations of Appendix F suggest is not needed; the trained models operate at $V/r = 16$, below the regime in which it has force. (iv) The window shares of Section 6 describe the power schedules (Remark 6.6). (v) The correlated commitment theorems (Section 4.5) need bounded conditional odds; hard constraints of language are covered only through the mixture form, whose context term must be controlled separately. (vi) The coarse-first ordering of the flow (Corollary 4.13) is proven for product-uniform data and exponent staggering; the flow is absorbed for almost every, not every, start; and the learning guarantee covers one-step sampling, not refinement; for correlated data the flow-level order is open (Remark 4.14).

Conclusion. On products of simplices the Beckmann field is explicit, its flow reaches the data for every schedule, and the time at which each block is decided is located by a non-asymptotic threshold with a universal Gaussian profile. Reading training time in the coordinate $\log(-\log(1 - t))$ turns the choice of schedule into a one-dimensional problem with an explicit answer, and this answer improves DBTM in training at no cost. A coarse-to-fine code gives no free signal; what it changes is the order in which a token is decided and the size of the state a network has to read. In our experiments the first shows up as a better coarse decision and a better use of context, and the remaining gap to flat DBTM sits in refinement. Testing the schedule rule and the coarse-to-fine state at language-model scale is the next step.

References

- [1] Manan Agarwal, Sheel Shah, Chanhyuk Lee, Jaehoon Yoo, Jerry Huang, Seunghoon Hong, Aditi Raghunathan, Jinwoo Kim, and Nicholas M. Boffi. Posterior refinement: Fast language generation via any-order flow maps. *arXiv preprint arXiv:2606.24773*, 2026.
- [2] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

- [3] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.
- [4] Luca Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking and critical instability. *Entropy*, 2025.
- [5] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021.
- [6] Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15, 2024.
- [7] Franco Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999.
- [8] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2024.
- [9] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems*, 2025.
- [10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- [11] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.
- [12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.
- [13] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning*, 2024.
- [14] Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İlkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. In *Advances in Neural Information Processing Systems*, 2024.
- [15] Bernard Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79, 1980.
- [16] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.
- [17] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025.
- [18] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, 2024.
- [19] Joshua Goodman. Classes for fast maximum entropy training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.
- [20] Robert D. Gordon. Values of Mills’ ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.
- [21] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018.

- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.
- [23] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
- [24] Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.
- [25] Cheuk-Kit Lee, Florentin Coeurdoux, Yuyuan Chen, Sophia Tang, Peter Potapchik, Yilun Du, Michael S. Albergo, and Eric Vanden-Eijnden. Beckmann transport models: From autonomous flows to one-step maps. *arXiv preprint arXiv:2608.01692*, 2026.
- [26] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [27] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.
- [28] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.
- [29] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024.
- [30] Andriy Mnih and Geoffrey E. Hinton. A scalable hierarchical distributed language model. In *Advances in Neural Information Processing Systems*, 2008.
- [31] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *International Conference on Artificial Intelligence and Statistics*, 2005.
- [32] Mitio Nagumo. Über die lage der integralkurven gewöhnlicher differentialgleichungen. *Proceedings of the Physico-Mathematical Society of Japan*, 24:551–559, 1942.
- [33] Mathias Ollu and Nikos Komodakis. Denoising hierarchical representations: Joint continuous diffusion for language modeling. *arXiv preprint arXiv:2610.08738*, 2026.
- [34] Emanuele Paolini and Eugene Stepanov. Decomposition of acyclic normal currents in a metric space. *Journal of Functional Analysis*, 263(11):3358–3390, 2012.
- [35] Peter Potapchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S. Albergo. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026.
- [36] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, 2023.
- [37] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, 2023.
- [38] Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, Ismail Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. In *International Conference on Machine Learning*, 2026.
- [39] Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

- [40] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1), 2025.
- [41] James Serrin and Dale E. Varberg. A general chain rule for derivatives and the change of variables formula for the Lebesgue integral. *The American Mathematical Monthly*, 76(5):514–520, 1969.
- [42] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, 2024.
- [43] Stanislav K. Smirnov. Decomposition of solenoidal vector charges into elementary solenoids, and the structure of normal one-dimensional flows. *St. Petersburg Mathematical Journal*, 5(4):841–867, 1994. Russian original: *Algebra i Analiz* 5(4):206–238, 1993.
- [44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [45] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.
- [46] Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. In *International Conference on Machine Learning*, 2024.
- [47] Sophia Tang and Shiyi Wang. Discrete beckmann transport models for one-step language modeling and reasoning. *arXiv preprint arXiv:2609.15903*, 2026.
- [48] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, 2024.
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- [50] Zihao Wu, Haoming Yang, Juncheng Dong, and Vahid Tarokh. Rethinking token prediction: Tree-structured diffusion language model. *arXiv preprint arXiv:2604.03537*, 2026.
- [51] Zhendong Yu and Haiping Huang. Nonequilibrium physics of generative diffusion models. *Physical Review E*, 111(1):014111, 2025.
- [52] Cai Zhou, Chenyu Wang, Dinghuai Zhang, Shangyuan Tong, Yifei Wang, Stephen Bates, and Tommi Jaakkola. Next semantic scale prediction via hierarchical diffusion language models. In *Advances in Neural Information Processing Systems*, 2025.

A Notation

symbol	meaning
V, L, K	vocabulary size, sequence length, depth of the code
$\gamma, c^{(k)}(w), v_k$	hierarchical code, level- k symbol of token w , number of symbols at level k
(ℓ, k)	block: position $\ell \in [L]$, level $k \in [K]$
$\mathbf{x} = (\mathbf{x}^{\ell,k}), \mathbf{x}^{\ell,k} \in \mathbb{R}^{v_k}$	state and its blocks; $D = L \sum_k v_k$
$\mathbf{e}_i \in \mathbb{R}^{v_k}$	i -th vertex (one-hot vector) of Δ^{v_k-1}
$\mathbf{j} = (j_{\ell,k}) \in \mathcal{J}, \mathbf{e}_j$	data configuration and its vertex in $\mathbb{R}^D$
$\mathcal{M}_1, \mathcal{P}$	vertex configurations; product of simplices $\prod_{\ell,k} \Delta^{v_k-1}$
$\mu_0 = \mathcal{N}(\mathbf{0}, \Sigma), \Sigma = \text{diag}(\sigma_k^2 \mathbf{I}_{v_k})$	prior
$\mu_1 = \sum_j \pi_j \delta_{\mathbf{e}_j}$	data distribution
$\alpha_k(t) = (1-t)^{a_k}, \mathbf{A}(t)$	level schedule and its block-diagonal matrix
$\mathbf{I}_t = \mathbf{A}(t)\mathbf{x}_0 + (\mathbf{I} - \mathbf{A}(t))\mathbf{x}_1$	coarse-to-fine interpolant
$p(\mathbf{x} t, \mathbf{j}), q(\mathbf{x}), w_j(t, \mathbf{x})$	likelihood, marginal density of $\mathbf{I}_t$, posterior over $(t, \mathbf{j})$
$\mathbf{b}, \mathbf{X}_s, \mathbf{T}, \mathbf{J}_T$	autonomous field, flow, transport map, Jacobian matrix of $\mathbf{T}$
$\rho_k(t) = (1 - \alpha_k(t))/(\alpha_k(t)\sigma_k)$	signal-to-noise ratio of level k
$\beta_k(t) = (1 - \alpha_k(t))/(\alpha_k(t)^2\sigma_k^2)$	inverse temperature of the level- k posterior
$\pi_*, s_n = \sqrt{2 \log(n-1)}$	posterior mass of the true vertex; threshold
$\varphi, \Phi, Q = 1 - \Phi$	standard normal density, distribution function, tail

B Proofs for Section 3

Throughout this section [Assumption 3.2](#) holds. We write $\tau := 1 - t \in (0, 1]$, so that $\alpha_k(t) = \tau^{a_k}$ and $\dot{\alpha}_k(t) = \frac{d}{dt}(1-t)^{a_k} = -a_k(1-t)^{a_k-1}$, hence

$$\frac{\dot{\alpha}_k(t)}{\alpha_k(t)} = \frac{-a_k(1-t)^{a_k-1}}{(1-t)^{a_k}} = -\frac{a_k}{1-t} \quad (t \in [0, 1)). \quad (37)$$

For $t \in [0, 1)$ the matrices $\mathbf{A}(t) = \text{diag}(\alpha_k(t)\mathbf{I}_{v_k})_{\ell,k}$ and $\dot{\mathbf{A}}(t) = \text{diag}(\dot{\alpha}_k(t)\mathbf{I}_{v_k})_{\ell,k}$ are block diagonal with scalar blocks; for $t \in [0, 1)$, $\mathbf{A}(t)$ is invertible with $\dot{\mathbf{A}}(t)\mathbf{A}(t)^{-1} = \text{diag}(\frac{\dot{\alpha}_k(t)}{\alpha_k(t)}\mathbf{I}_{v_k})_{\ell,k}$.

B.1 The law of the interpolant

Lemma B.1 (Likelihood). *For $t \in (0, 1)$ and $\mathbf{j} \in \mathcal{J}$, conditionally on t and $\mathbf{x}_1 = \mathbf{e}_j$ the interpolant has the Gaussian law $\mathcal{N}((\mathbf{I} - \mathbf{A}(t))\mathbf{e}_j, \mathbf{A}(t)\Sigma\mathbf{A}(t))$, whose density is [Eq. \(4\)](#).*

Proof. Idea. Given t and the endpoint, the interpolant is an affine image of the Gaussian prior sample, so its law is Gaussian and only the mean and the covariance have to be read off. Because $\mathbf{A}(t)$ and Σ are diagonal with scalar blocks, the covariance is diagonal and the density splits into a product over blocks, which is the form of [Eq. \(4\)](#).

Conditionally on t and $\mathbf{x}_1 = \mathbf{e}_j$, $\mathbf{I}_t = \mathbf{A}(t)\mathbf{x}_0 + (\mathbf{I} - \mathbf{A}(t))\mathbf{e}_j$ is an affine function of $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \Sigma)$, with $\mathbf{x}_0$ independent of $(t, \mathbf{x}_1)$. An affine map $\mathbf{z} \mapsto \mathbf{M}\mathbf{z} + \mathbf{m}$ sends $\mathcal{N}(\mathbf{0}, \Sigma)$ to $\mathcal{N}(\mathbf{m}, \mathbf{M}\Sigma\mathbf{M}^\top)$. Here $\mathbf{M} = \mathbf{A}(t)$, which is symmetric, and $\mathbf{m} = (\mathbf{I} - \mathbf{A}(t))\mathbf{e}_j$, so the conditional law is $\mathcal{N}((\mathbf{I} - \mathbf{A})\mathbf{e}_j, \mathbf{A}\Sigma\mathbf{A})$. Since $\mathbf{A}$ and Σ are diagonal with scalar blocks,

$$\Sigma_t := \mathbf{A}(t)\Sigma\mathbf{A}(t) = \text{diag}(\alpha_k(t)^2\sigma_k^2\mathbf{I}_{v_k})_{\ell,k}, \quad \det \Sigma_t = \prod_{\ell,k} (\alpha_k(t)^2\sigma_k^2)^{v_k} > 0,$$

because $\alpha_k(t) = (1-t)^{a_k} > 0$ for $t \in (0, 1)$. Hence the law has the density $(2\pi)^{-D/2}(\det \Sigma_t)^{-1/2} \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \Sigma_t^{-1}(\mathbf{x} - \mathbf{m}))$. Block (ℓ, k) of $\mathbf{m}$ is $(1 - \alpha_k)\mathbf{e}_{j_{\ell,k}}$, so

$$(\mathbf{x} - \mathbf{m})^\top \Sigma_t^{-1}(\mathbf{x} - \mathbf{m}) = \sum_{\ell,k} \frac{\|\mathbf{x}^{\ell,k} - (1 - \alpha_k)\mathbf{e}_{j_{\ell,k}}\|^2}{\alpha_k^2\sigma_k^2}, \quad (2\pi)^{-D/2}(\det \Sigma_t)^{-1/2} = \prod_{\ell,k} (2\pi\alpha_k^2\sigma_k^2)^{-v_k/2},$$

using $D = L \sum_k v_k = \sum_{\ell,k} v_k$. Inserting both identities gives [Eq. \(4\)](#). $\square$

Lemma B.2 (Velocity on a fibre). *For $t \in (0, 1)$, on the event $\{\mathbf{I}_t = \mathbf{x}, \mathbf{x}_1 = \mathbf{e}_j\}$ the velocity is*

$$\dot{\mathbf{I}}_t = \dot{\mathbf{A}}(t) \mathbf{A}(t)^{-1} (\mathbf{x} - \mathbf{e}_j), \quad \text{i.e.} \quad \dot{\mathbf{I}}_t^{\ell,k} = \frac{\dot{\alpha}_k(t)}{\alpha_k(t)} (\mathbf{x}^{\ell,k} - \mathbf{e}_{j_{\ell,k}}) = \frac{a_k}{1-t} (\mathbf{e}_{j_{\ell,k}} - \mathbf{x}^{\ell,k}). \quad (38)$$

Proof. Idea. The velocity $\dot{\mathbf{A}}(t)(\mathbf{x}_0 - \mathbf{x}_1)$ involves the prior sample, which is not observed; what must be shown is that on the fibre $\{\mathbf{I}_t = \mathbf{x}, \mathbf{x}_1 = \mathbf{e}_j\}$ it is a function of $\mathbf{x}$ alone. Since $\mathbf{A}(t)$ is invertible for $t < 1$, [Eq. \(3\)](#) can be solved for $\mathbf{x}_0$, and substituting the solution gives the formula.

Differentiating [Eq. \(3\)](#) in t with $\mathbf{x}_0, \mathbf{x}_1$ fixed gives $\dot{\mathbf{I}}_t = \dot{\mathbf{A}}(t) \mathbf{x}_0 - \dot{\mathbf{A}}(t) \mathbf{x}_1 = \dot{\mathbf{A}}(t)(\mathbf{x}_0 - \mathbf{x}_1)$. On $\{\mathbf{I}_t = \mathbf{x}, \mathbf{x}_1 = \mathbf{e}_j\}$ we can solve [Eq. \(3\)](#) for the prior sample, since $\mathbf{A}(t)$ is invertible for $t \in (0, 1)$:

$$\mathbf{x}_0 = \mathbf{A}(t)^{-1} (\mathbf{x} - (\mathbf{I} - \mathbf{A}(t)) \mathbf{e}_j) = \mathbf{A}(t)^{-1} \mathbf{x} - \mathbf{A}(t)^{-1} \mathbf{e}_j + \mathbf{e}_j.$$

Therefore $\mathbf{x}_0 - \mathbf{x}_1 = \mathbf{x}_0 - \mathbf{e}_j = \mathbf{A}(t)^{-1} \mathbf{x} - \mathbf{A}(t)^{-1} \mathbf{e}_j = \mathbf{A}(t)^{-1} (\mathbf{x} - \mathbf{e}_j)$, and $\dot{\mathbf{I}}_t = \dot{\mathbf{A}}(t) \mathbf{A}(t)^{-1} (\mathbf{x} - \mathbf{e}_j)$. Both matrices are block diagonal with scalar blocks, so block (ℓ, k) equals $(\dot{\alpha}_k / \alpha_k) (\mathbf{x}^{\ell,k} - \mathbf{e}_{j_{\ell,k}})$, and [Eq. \(37\)](#) gives the last expression. $\square$

B.2 Integrability and smoothness

Lemma B.3 (Integrability). (i) $\mathbb{E} \|\dot{\mathbf{I}}_t\| < \infty$, so the conditional expectation [Eq. \(1\)](#) is well defined.

(ii) For every $\mathbf{x} \in \mathbb{R}^D$, $q(\mathbf{x}) > 0$, and for every $\mathbf{x} \notin \mathcal{M}_1$, $q(\mathbf{x}) < \infty$.

(iii) For every $\mathbf{x} \notin \mathcal{M}_1$, every $\mathbf{j}$ and every k , $\int_0^1 p(\mathbf{x} | t, \mathbf{j}) \frac{a_k}{1-t} dt < \infty$.

(iv) For every $\bar{\mathbf{x}} \notin \mathcal{M}_1$ there are a neighbourhood $U \ni \bar{\mathbf{x}}$ and, for every multi-index of derivatives β , a function $G_\beta \in L^1(0, 1)$ such that $\sup_{\mathbf{x} \in U} |\partial_{\mathbf{x}}^\beta p(\mathbf{x} | t, \mathbf{j})| (1 + \frac{a_k}{1-t})(1 + \|\mathbf{x}\|) \leq G_\beta(t)$ for all $t \in (0, 1)$, $\mathbf{j} \in \mathcal{J}$, $k \in [K]$.

Proof. Idea. Apart from (i), the only danger is the endpoint $t \rightarrow 1$, where the prefactor of [Eq. \(4\)](#) grows like τ^{-N} and the speed $a_k/(1-t)$ like τ^{-1} . At a point $\mathbf{x}$ at distance $d > 0$ from the vertex $\mathbf{e}_j$, however, the Gaussian of configuration $\mathbf{j}$ concentrates at scale $\alpha_k = \tau^{a_k}$ around a point close to $\mathbf{e}_j$, so $p(\mathbf{x} | t, \mathbf{j})$ carries a factor $\exp(-c\tau^{-2a_{\min}})$, which beats every power of τ ; this one estimate gives (iii), and (ii) follows from it. Part (iv) repeats the estimate uniformly on a ball around $\bar{\mathbf{x}}$ and for $\mathbf{x}$ -derivatives, which are p times polynomials in quantities of size at most $\tau^{-2a_{\max}}$ and so are still dominated by the same exponential. Part (i) needs no such estimate: it follows from $\int_0^1 |\dot{\alpha}_k| dt = 1$ and the Gaussian moments of $\mathbf{x}_0$.

Throughout, $\kappa_0 := \prod_{\ell,k} (2\pi\sigma_k^2)^{-v_k/2}$, $a_{\min} = \min_k a_k$, $a_{\max} = \max_k a_k$, $\sigma_{\min} = \min_k \sigma_k$, $\sigma_{\max} = \max_k \sigma_k$, and

$$\prod_{\ell,k} (2\pi\alpha_k(t)^2\sigma_k^2)^{-v_k/2} = \kappa_0 \prod_{\ell,k} \alpha_k(t)^{-v_k} = \kappa_0 \tau^{-\sum_{\ell,k} a_k v_k} = \kappa_0 \tau^{-N}, \quad N := L \sum_k a_k v_k. \quad (39)$$

(i) The event $\{t = 1\}$ has probability zero, and for $t \in [0, 1)$ the proof of [Lemma B.2](#) gives $\dot{\mathbf{I}}_t = \dot{\mathbf{A}}(t)(\mathbf{x}_0 - \mathbf{x}_1)$. Blockwise, using $\|\mathbf{x}_1^{\ell,k}\| = \|\mathbf{e}_{j_{\ell,k}}\| = 1$,

$$\|\dot{\mathbf{I}}_t^{\ell,k}\| = |\dot{\alpha}_k(t)| \|\mathbf{x}_0^{\ell,k} - \mathbf{x}_1^{\ell,k}\| \leq |\dot{\alpha}_k(t)| (\|\mathbf{x}_0^{\ell,k}\| + 1).$$

The Euclidean norm of a vector is at most the sum of the norms of its blocks, and t is independent of $\mathbf{x}_0$, so

$$\mathbb{E} \|\dot{\mathbf{I}}_t\| \leq \sum_{\ell,k} \mathbb{E} \|\dot{\mathbf{I}}_t^{\ell,k}\| \leq \sum_{\ell,k} \mathbb{E} |\dot{\alpha}_k(t)| \mathbb{E} (\|\mathbf{x}_0^{\ell,k}\| + 1) = \sum_{\ell,k} \left(\int_0^1 |\dot{\alpha}_k(t)| dt \right) (\mathbb{E} \|\mathbf{x}_0^{\ell,k}\| + 1).$$

By monotone convergence, $\int_0^1 |\dot{\alpha}_k| dt = \lim_{t' \uparrow 1} \int_0^{t'} a_k (1-t)^{a_k-1} dt = \lim_{t' \uparrow 1} (1 - (1-t')^{a_k}) = 1$ (for $a_k < 1$ the integrand is unbounded near $t = 1$ but integrable). By the Cauchy–Schwarz inequality and $\mathbf{x}_0^{\ell,k} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}_{v_k})$, $\mathbb{E} \|\mathbf{x}_0^{\ell,k}\| \leq (\mathbb{E} \|\mathbf{x}_0^{\ell,k}\|^2)^{1/2} = \sigma_k \sqrt{v_k}$. Hence $\mathbb{E} \|\dot{\mathbf{I}}_t\| \leq \sum_{\ell,k} (\sigma_k \sqrt{v_k} + 1) < \infty$.

(iii) Fix $\mathbf{x} \notin \mathcal{M}_1$ and $\mathbf{j}$. Since $\mathbf{x} \neq \mathbf{e}_{\mathbf{j}}$, there is a block (ℓ^*, k^*) with $d := \|\mathbf{x}^{\ell^*, k^*} - \mathbf{e}_{j_{\ell^*, k^*}}\| > 0$. By the triangle inequality and $\|\mathbf{e}_{j_{\ell^*, k^*}}\| = 1$,

$$\|\mathbf{x}^{\ell^*, k^*} - (1 - \alpha_{k^*}) \mathbf{e}_{j_{\ell^*, k^*}}\| = \|(\mathbf{x}^{\ell^*, k^*} - \mathbf{e}_{j_{\ell^*, k^*}}) + \alpha_{k^*} \mathbf{e}_{j_{\ell^*, k^*}}\| \geq d - \alpha_{k^*}.$$

Let $\tau_0 := \min_k (d/2)^{1/a_k} \wedge 1$. For $\tau \leq \tau_0$ we have $\alpha_{k^*} = \tau^{a_{k^*}} \leq ((d/2)^{1/a_{k^*}})^{a_{k^*}} = d/2$, so the norm above is at least $d/2$ and the factor of block (ℓ^*, k^*) in Eq. (4) is at most $\exp(-d^2/(8\alpha_{k^*}^2 \sigma_{k^*}^2))$. Bounding every other exponential factor in Eq. (4) by 1 and using Eq. (39),

$$p(\mathbf{x} | t, \mathbf{j}) \leq \kappa_0 \tau^{-N} \exp\left(-\frac{d^2}{8\sigma_{k^*}^2} \tau^{-2a_{k^*}}\right) \quad (\tau \leq \tau_0).$$

Since $\tau \leq 1$ and $a_{k^*} \geq a_{\min}$, $\tau^{-2a_{k^*}} \geq \tau^{-2a_{\min}}$; also $\sigma_{k^*} \leq \sigma_{\max}$. Multiplying by $a_k/(1-t) = a_k \tau^{-1}$,

$$p(\mathbf{x} | t, \mathbf{j}) \frac{a_k}{1-t} \leq C \tau^{-N-1} \exp(-c \tau^{-2a_{\min}}), \quad C = a_k \prod_{\ell, k'} (2\pi \sigma_{k'}^2)^{-v_{k'}/2}, \quad c = \frac{d^2}{8\sigma_{\max}^2}, \quad (40)$$

for $\tau \leq \tau_0$. With $y = 1/\tau \in [1/\tau_0, \infty)$ the right-hand side is $h(y) := C y^{N+1} e^{-cy^{2a_{\min}}}$, a continuous function with $h(y) \rightarrow 0$ as $y \rightarrow \infty$ (the exponential decays faster than any power), hence bounded on $[1/\tau_0, \infty)$. So the integrand is bounded for $t \in [1 - \tau_0, 1)$. For $t \in [0, 1 - \tau_0]$ we have $\tau \geq \tau_0$, so $\alpha_{k'} \geq \tau_0^{a_{k'}}$ for every k' ; bounding the exponentials by 1, $p(\mathbf{x} | t, \mathbf{j}) \leq \prod_{\ell, k'} (2\pi \tau_0^{2a_{k'}} \sigma_{k'}^2)^{-v_{k'}/2}$, and $a_k/(1-t) \leq a_k/\tau_0$. The integrand is continuous on $[0, 1)$ and bounded, hence integrable on $(0, 1)$.

(ii) For fixed $\mathbf{x}$ and $t \in (0, 1)$, every $p(\mathbf{x} | t, \mathbf{j})$ is a value of a Gaussian density with nonsingular covariance (Lemma B.1), hence strictly positive, and $\sum_{\mathbf{j}} \pi_{\mathbf{j}} = 1$; so $t \mapsto \sum_{\mathbf{j}} \pi_{\mathbf{j}} p(\mathbf{x} | t, \mathbf{j})$ is continuous and strictly positive on $(0, 1)$, and its integral $q(\mathbf{x})$ is strictly positive. For $\mathbf{x} \notin \mathcal{M}_1$, $1-t \leq 1$ gives $p(\mathbf{x} | t, \mathbf{j}) \leq \frac{1}{a_1} p(\mathbf{x} | t, \mathbf{j}) \frac{a_1}{1-t}$, so by (iii) (with $k = 1$)

$$q(\mathbf{x}) = \sum_{\mathbf{j}} \pi_{\mathbf{j}} \int_0^1 p(\mathbf{x} | t, \mathbf{j}) dt \leq \frac{1}{a_1} \sum_{\mathbf{j}} \pi_{\mathbf{j}} \int_0^1 p(\mathbf{x} | t, \mathbf{j}) \frac{a_1}{1-t} dt < \infty.$$

For $\mathbf{x} = \mathbf{e}_{\mathbf{j}'} \in \mathcal{M}_1$ finiteness can fail: since $\mathbf{x}^{\ell, k} - (1 - \alpha_k) \mathbf{e}_{j'_{\ell, k}} = \alpha_k \mathbf{e}_{j'_{\ell, k}}$, the term $\mathbf{j} = \mathbf{j}'$ equals $\pi_{\mathbf{j}'} \kappa_0 \tau^{-N} \exp(-L \sum_k 1/(2\sigma_k^2))$ by Eq. (39), whose integral over $\tau \in (0, 1)$ is infinite when $\pi_{\mathbf{j}'} > 0$ and $N \geq 1$. We therefore only claim $q(\mathbf{x}) < \infty$ for $\mathbf{x} \notin \mathcal{M}_1$, which is all we use ($\mathcal{M}_1$ is a finite, hence Lebesgue-null, set).

(iv) Let $d := \min_{\mathbf{j}} \max_{\ell, k} \|\bar{\mathbf{x}}^{\ell, k} - \mathbf{e}_{j_{\ell, k}}\|$. For each $\mathbf{j}$, $\bar{\mathbf{x}} \neq \mathbf{e}_{\mathbf{j}}$, so the maximum is positive; d is a minimum of finitely many positive numbers, so $d > 0$. Let U be the open ball of radius $d/4$ around $\bar{\mathbf{x}}$. For $\mathbf{x} \in U$ and every $\mathbf{j}$, let (ℓ, k) be a block attaining the maximum for $\bar{\mathbf{x}}$; then $\|\mathbf{x}^{\ell, k} - \mathbf{e}_{j_{\ell, k}}\| \geq \|\bar{\mathbf{x}}^{\ell, k} - \mathbf{e}_{j_{\ell, k}}\| - \|\mathbf{x} - \bar{\mathbf{x}}\| \geq d - d/4 = 3d/4$. The argument of (iii), with d replaced by $3d/4$, therefore holds uniformly in $\mathbf{x} \in U$ and $\mathbf{j}$: with $\tau'_0 := \min_k (3d/8)^{1/a_k} \wedge 1$ and $c' := 9d^2/(128\sigma_{\max}^2)$,

$$p(\mathbf{x} | t, \mathbf{j}) \leq \kappa_0 \tau^{-N} \exp(-c' \tau^{-2a_{\min}}) \quad (\tau \leq \tau'_0), \quad p(\mathbf{x} | t, \mathbf{j}) \leq \prod_{\ell, k} (2\pi \tau_0'^{2a_k} \sigma_k^2)^{-v_k/2} \quad (\tau \geq \tau'_0), \quad (41)$$

for all $\mathbf{x} \in U$ and $\mathbf{j} \in \mathcal{J}$.

Derivatives. Fix t and $\mathbf{j}$, and set $s_k := 1/(\alpha_k^2 \sigma_k^2)$ and $u_i^{\ell, k} := s_k (x_i^{\ell, k} - (1 - \alpha_k)(\mathbf{e}_{j_{\ell, k}})_i)$. By Eq. (4), $\log p = \text{const}(t) - \frac{1}{2} \sum_{\ell, k} s_k \|\mathbf{x}^{\ell, k} - (1 - \alpha_k) \mathbf{e}_{j_{\ell, k}}\|^2$, so

$$\partial_{x_i^{\ell, k}} p = -u_i^{\ell, k} p, \quad \partial_{x_i^{\ell, k}} u_{i'}^{\ell', k'} = s_k \delta_{(\ell, k, i), (\ell', k', i')}, \quad \partial_{x_i^{\ell, k}} s_{k'} = 0.$$

By induction on $|\beta|$, $\partial_{\mathbf{x}}^{\beta} p = p P_{\beta}(\mathbf{u}, \mathbf{s})$ with a polynomial P_{β} of degree at most $|\beta|$ whose coefficients depend only on β : $P_0 = 1$, and $\partial_{x_i^{\ell, k}} (p P) = p (-u_i^{\ell, k} P + \partial_{x_i^{\ell, k}} P)$, where $-u_i^{\ell, k} P$ raises the degree by one and $\partial_{x_i^{\ell, k}} P$

(which replaces a factor $u_i^{\ell,k}$ by s_k) does not raise it. On U and for $\tau \leq 1$, with $R_U := \|\bar{\mathbf{x}}\| + d/4$,

$$|u_i^{\ell,k}| \leq \frac{\|\mathbf{x}^{\ell,k}\| + 1}{\sigma_k^2 \tau^{2a_k}} \leq \frac{R_U + 1}{\sigma_{\min}^2} \tau^{-2a_{\max}}, \quad s_k \leq \frac{1}{\sigma_{\min}^2} \tau^{-2a_{\max}}.$$

With $\Lambda := \max\{1, (R_U + 1)/\sigma_{\min}^2\}$, each monomial of degree $m \leq |\beta|$ is at most $\Lambda^m \tau^{-2a_{\max}m} \leq \Lambda^{|\beta|} \tau^{-2a_{\max}|\beta|}$, so $|P_\beta| \leq C' \tau^{-N'}$ with $N' := 2a_{\max}|\beta|$ and C' equal to $\Lambda^{|\beta|}$ times the sum of the absolute values of the coefficients of P_β . Finally $(1 + \|\mathbf{x}\|) \leq 1 + R_U$ on U and $1 + a_k/(1 - t) \leq (1 + a_{\max})\tau^{-1}$ for $\tau \leq 1$.

The dominating function. For $\tau \leq \tau'_0$, by Eq. (41),

$$|\partial_{\mathbf{x}}^\beta p| \left(1 + \frac{a_k}{1 - t}\right) (1 + \|\mathbf{x}\|) \leq C'' \tau^{-N - N' - 1} \exp(-c' \tau^{-2a_{\min}}), \quad C'' := \kappa_0 C' (1 + a_{\max})(1 + R_U),$$

which is bounded in $\tau \in (0, \tau'_0]$ by the argument after Eq. (40). For $\tau \geq \tau'_0$ the left-hand side is at most $\prod_{\ell,k} (2\pi \tau_0'^{2a_k} \sigma_k^2)^{-v_k/2} C' \tau_0'^{-N'} (1 + a_{\max}/\tau'_0) (1 + R_U)$. Both bounds are independent of $\mathbf{x} \in U$, $\mathbf{j}$ and k , so their maximum is a constant G_β , which lies in $L^1(0, 1)$. $\square$

B.3 Proof of Proposition 3.3

Proof. Idea. To identify $\mathbb{E}[\dot{\mathbf{I}}_t \mid \mathbf{I}_t]$ it suffices to find a function $\mathbf{m}$ with $\mathbb{E}[\dot{\mathbf{I}}_t \phi(\mathbf{I}_t)] = \mathbb{E}[\phi(\mathbf{I}_t) \mathbf{m}(\mathbf{I}_t)]$ for every bounded measurable ϕ . Conditionally on t and the endpoint, both the law of $\mathbf{I}_t$ (Lemma B.1) and the velocity as a function of $\mathbf{I}_t$ (Lemma B.2) are explicit, so we disintegrate over $(t, \mathbf{x}_1)$ and then exchange the order of integration; Lemma B.3(i) is what makes the exchange legitimate. Smoothness is differentiation under the integral sign in t , with the dominating functions of Lemma B.3(iv).

Disintegration. By independence of t , $\mathbf{x}_0$, $\mathbf{x}_1$, for every Borel $F : [0, 1) \times \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, \infty]$ (Tonelli), and for every F with $\mathbb{E}|F(t, \mathbf{x}_0, \mathbf{x}_1)| < \infty$ (Fubini),

$$\mathbb{E}F(t, \mathbf{x}_0, \mathbf{x}_1) = \int_0^1 \sum_j \pi_j \int_{\mathbb{R}^D} F(t, \mathbf{z}, \mathbf{e}_j) \mathcal{N}(\mathbf{0}, \Sigma) d\mathbf{z} dt. \quad (42)$$

If, for $t \in (0, 1)$ fixed, $F(t, \mathbf{z}, \mathbf{e}_j) = f_{t,j}(\mathbf{A}(t)\mathbf{z} + (\mathbf{I} - \mathbf{A}(t))\mathbf{e}_j)$, then by Lemma B.1 the inner integral equals $\int f_{t,j}(\mathbf{x}) p(\mathbf{x} \mid t, \mathbf{j}) d\mathbf{x}$ (image of a measure under the affine map). By Lemma B.2, $\dot{\mathbf{I}}_t = \dot{\mathbf{A}}(t)\mathbf{A}(t)^{-1}(\mathbf{I}_t - \mathbf{e}_j)$ when $\mathbf{x}_1 = \mathbf{e}_j$, so this applies to $F = \dot{\mathbf{I}}_t \phi(\mathbf{I}_t)$, coordinatewise, with

$$f_{t,j}(\mathbf{x}) = \mathbf{v}_{t,j}(\mathbf{x}) \phi(\mathbf{x}), \quad \mathbf{v}_{t,j}(\mathbf{x}) := \dot{\mathbf{A}}(t)\mathbf{A}(t)^{-1}(\mathbf{x} - \mathbf{e}_j).$$

Marginal density. Taking $F = \mathbb{1}_B(\mathbf{I}_t)$ for a Borel set B in Eq. (42) and exchanging the integrals by Tonelli,

$$\mathbb{P}(\mathbf{I}_t \in B) = \int_0^1 \sum_j \pi_j \int_B p(\mathbf{x} \mid t, \mathbf{j}) d\mathbf{x} dt = \int_B q(\mathbf{x}) d\mathbf{x},$$

so q is the density of $\mathbf{I}_t$.

The field. Let $\phi : \mathbb{R}^D \rightarrow \mathbb{R}$ be bounded and measurable. Applying Eq. (42) to $F = \|\dot{\mathbf{I}}_t\|$, the function $(t, \mathbf{j}, \mathbf{x}) \mapsto \pi_j \|\mathbf{v}_{t,j}(\mathbf{x})\| p(\mathbf{x} \mid t, \mathbf{j})$ has integral $\mathbb{E}\|\dot{\mathbf{I}}_t\| < \infty$ over $(0, 1) \times \mathcal{J} \times \mathbb{R}^D$ by Lemma B.3(i). The integrand of the next display has norm at most $\|\phi\|_\infty$ times this function, so it is absolutely integrable, and Fubini's theorem allows us to integrate over t and $\mathbf{j}$ first:

$$\begin{aligned} \mathbb{E}[\dot{\mathbf{I}}_t \phi(\mathbf{I}_t)] &= \int_0^1 \sum_j \pi_j \int_{\mathbb{R}^D} \mathbf{v}_{t,j}(\mathbf{x}) \phi(\mathbf{x}) p(\mathbf{x} \mid t, \mathbf{j}) d\mathbf{x} dt = \int_{\mathbb{R}^D} \phi(\mathbf{x}) \mathbf{n}(\mathbf{x}) d\mathbf{x}, \\ \mathbf{n}(\mathbf{x}) &:= \int_0^1 \sum_j \pi_j p(\mathbf{x} \mid t, \mathbf{j}) \mathbf{v}_{t,j}(\mathbf{x}) dt. \end{aligned}$$

For $\mathbf{x} \notin \mathcal{M}_1$ the integral $\mathbf{n}(\mathbf{x})$ converges absolutely: by Eq. (38), $\|\mathbf{v}_{t,j}^{\ell,k}(\mathbf{x})\| = \frac{a_k}{1-t} \|\mathbf{e}_{j_{\ell,k}} - \mathbf{x}^{\ell,k}\| \leq \frac{a_k}{1-t} (1 + \|\mathbf{x}\|)$, and $\int_0^1 p(\mathbf{x} \mid t, \mathbf{j}) \frac{a_k}{1-t} dt < \infty$ by Lemma B.3(iii). By Lemma B.3(ii), $0 < q(\mathbf{x}) < \infty$ there, so

$$\mathbf{m}(\mathbf{x}) := \frac{\mathbf{n}(\mathbf{x})}{q(\mathbf{x})} = \int_0^1 \sum_j w_j(t, \mathbf{x}) \dot{\mathbf{A}}(t)\mathbf{A}(t)^{-1}(\mathbf{x} - \mathbf{e}_j) dt \quad (\mathbf{x} \notin \mathcal{M}_1)$$

is well defined and finite (we use $w_j = \pi_j p/q$), and Borel (partial integrals of jointly measurable functions are measurable). Since $\mathcal{M}_1$ is Lebesgue-null,

$$\mathbb{E}[\dot{\mathbf{I}}_t \phi(\mathbf{I}_t)] = \int_{\mathbb{R}^D \setminus \mathcal{M}_1} \phi(\mathbf{x}) \mathbf{m}(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\phi(\mathbf{I}_t) \mathbf{m}(\mathbf{I}_t)].$$

Moreover $\mathbb{E}\|\mathbf{m}(\mathbf{I}_t)\| = \int \|\mathbf{n}\| d\mathbf{x} \leq \mathbb{E}\|\dot{\mathbf{I}}_t\| < \infty$. Since the identity holds for every bounded measurable ϕ and $\mathbf{m}(\mathbf{I}_t)$ is integrable and $\sigma(\mathbf{I}_t)$ -measurable, $\mathbf{m}(\mathbf{I}_t)$ is a version of $\mathbb{E}[\dot{\mathbf{I}}_t | \mathbf{I}_t]$. Thus $\mathbf{b} = \mathbf{m} q d\mathbf{x}$ -almost everywhere, i.e. Lebesgue-almost everywhere because $q > 0$; we take the everywhere-defined $\mathbf{m}$ as the definition of $\mathbf{b}$ off $\mathcal{M}_1$. Block (ℓ, k) of $\dot{\mathbf{A}}\mathbf{A}^{-1}(\mathbf{x} - \mathbf{e}_j)$ is $\frac{a_k}{1-t}(\mathbf{e}_{j_{\ell,k}} - \mathbf{x}^{\ell,k})$ by Eq. (38), which gives Eq. (5), and the integrals in Eq. (5) converge absolutely by the bound on $\mathbf{n}$ above.

Smoothness. We use the following standard fact. Let $U \subset \mathbb{R}^D$ be an open ball and $f : U \times (0, 1) \rightarrow \mathbb{R}$ be such that $f(\cdot, t) \in C^\infty(U)$ for every t , every $\partial_{\mathbf{x}}^\beta f$ is jointly measurable, and $\sup_{\mathbf{x} \in U} |\partial_{\mathbf{x}}^\beta f(\mathbf{x}, t)| \leq H_\beta(t)$ with $H_\beta \in L^1(0, 1)$. Then $F(\mathbf{x}) = \int_0^1 f(\mathbf{x}, t) dt$ is C^∞ on U and $\partial_{\mathbf{x}}^\beta F = \int_0^1 \partial_{\mathbf{x}}^\beta f dt$. Indeed, for a coordinate direction $\mathbf{e}$ and $\mathbf{x}, \mathbf{x} + h\mathbf{e} \in U$, the mean value theorem on the segment (contained in the convex set U) gives $|f(\mathbf{x} + h\mathbf{e}, t) - f(\mathbf{x}, t)|/|h| \leq H_{\mathbf{e}}(t)$, so dominated convergence yields $\partial_{\mathbf{e}} F = \int \partial_{\mathbf{e}} f dt$; continuity of $\int \partial_{\mathbf{x}}^\beta f dt$ in $\mathbf{x}$ follows by dominated convergence with H_β , and induction on $|\beta|$ gives the claim.

Fix $\bar{\mathbf{x}} \notin \mathcal{M}_1$ and take U and G_β from Lemma B.3(iv). The function $p(\mathbf{x} | t, \mathbf{j})$ is C^∞ in $\mathbf{x}$ and continuous in $(\mathbf{x}, t) \in \mathbb{R}^D \times (0, 1)$, as are its $\mathbf{x}$ -derivatives. For q , the integrand $\sum_j \pi_j p$ has $\mathbf{x}$ -derivatives bounded on U by G_β (the factors $1 + a_k/(1-t)$ and $1 + \|\mathbf{x}\|$ are at least 1). For coordinate i of block (ℓ, k) of $\mathbf{n}$, the integrand is $\sum_j \pi_j \frac{a_k}{1-t} p((\mathbf{e}_{j_{\ell,k}})_i - x_i^{\ell,k})$; by the Leibniz rule,

$$\partial_{\mathbf{x}}^\beta \left[p((\mathbf{e}_{j_{\ell,k}})_i - x_i^{\ell,k}) \right] = ((\mathbf{e}_{j_{\ell,k}})_i - x_i^{\ell,k}) \partial_{\mathbf{x}}^\beta p - \beta_{\ell,k,i} \partial_{\mathbf{x}}^{\beta - \mathbf{e}_{\ell,k,i}} p,$$

where $\beta_{\ell,k,i}$ is the order of differentiation in $x_i^{\ell,k}$ and the second term is absent if $\beta_{\ell,k,i} = 0$. Since $|(\mathbf{e}_{j_{\ell,k}})_i - x_i^{\ell,k}| \leq 1 + \|\mathbf{x}\|$, the β -th derivative of the integrand is bounded on U by $(1 + |\beta|) \sum_{\beta' \leq \beta} G_{\beta'}(t) \in L^1(0, 1)$. By the fact above, $\mathbf{n}$ and q are C^∞ on U . Since $q > 0$ on U (Lemma B.3(ii)), $\mathbf{b} = \mathbf{n}/q$ is C^∞ on U . As $\bar{\mathbf{x}}$ was arbitrary, $\mathbf{b} \in C^\infty(\mathbb{R}^D \setminus \mathcal{M}_1)$. $\square$

B.4 Proof of Proposition 3.4

Proof. Idea. Part (i) is a regrouping of Eq. (5) by the value of block (ℓ, k) of the endpoint, and the rest follows from it. Summing the coordinates of (i) shows that $h^{\ell,k} = \mathbf{1}^\top \mathbf{x}^{\ell,k} - 1$ obeys a scalar linear equation with a nonnegative coefficient, which an integrating factor solves in closed form; the same device applied to a single coordinate shows that a coordinate cannot become negative. This makes the forward invariance of $\mathcal{P}$ explicit, so Nagumo's theorem is cited only as the general principle behind the computation.

(i) Group the configurations in Eq. (5) by the value $i = j_{\ell,k}$ of block (ℓ, k) :

$$\mathbf{b}^{\ell,k}(\mathbf{x}) = \int_0^1 \sum_{i=1}^{v_k} \sum_{\mathbf{j}: j_{\ell,k}=i} w_{\mathbf{j}}(t, \mathbf{x}) \frac{a_k}{1-t} (\mathbf{e}_i - \mathbf{x}^{\ell,k}) dt = \sum_{i=1}^{v_k} \left(\int_0^1 \sum_{\mathbf{j}: j_{\ell,k}=i} w_{\mathbf{j}}(t, \mathbf{x}) \frac{a_k}{1-t} dt \right) (\mathbf{e}_i - \mathbf{x}^{\ell,k}).$$

Exchanging the finite sums with the integral is allowed because each $t \mapsto w_{\mathbf{j}}(t, \mathbf{x}) \frac{a_k}{1-t}$ is integrable for $\mathbf{x} \notin \mathcal{M}_1$ (Lemma B.3(ii)–(iii)), and $\mathbf{e}_i - \mathbf{x}^{\ell,k}$ does not depend on t . The coefficient in parentheses is $c_i^{\ell,k}(\mathbf{x})$, and it is nonnegative because $w_{\mathbf{j}} \geq 0$ and $a_k/(1-t) > 0$.

Continuity of the coefficients. For $\mathbf{x} \notin \mathcal{M}_1$, $c_i^{\ell,k}(\mathbf{x}) = q(\mathbf{x})^{-1} \sum_{\mathbf{j}: j_{\ell,k}=i} \pi_{\mathbf{j}} \int_0^1 p(\mathbf{x} | t, \mathbf{j}) \frac{a_k}{1-t} dt$. Near any $\bar{\mathbf{x}} \notin \mathcal{M}_1$ the integrand is continuous in $\mathbf{x}$ and dominated by G_0 of Lemma B.3(iv), so the integral is continuous by dominated convergence, and q is continuous and positive (proof of Proposition 3.3). Hence $c_i^{\ell,k}$ and $C^{\ell,k} = \sum_i c_i^{\ell,k}$ are continuous on $\mathbb{R}^D \setminus \mathcal{M}_1$.

(ii) Let $\mathbf{X}_s$ be a solution in $\mathbb{R}^D \setminus \mathcal{M}_1$; it is C^1 because $\mathbf{b}$ is continuous there. Using (i) and $\mathbf{1}^\top \mathbf{e}_i = 1$,

$$\dot{h}^{\ell,k}(s) = \mathbf{1}^\top \dot{\mathbf{X}}_s^{\ell,k} = \mathbf{1}^\top \mathbf{b}^{\ell,k}(\mathbf{X}_s) = \sum_i c_i^{\ell,k}(\mathbf{X}_s) (1 - \mathbf{1}^\top \mathbf{X}_s^{\ell,k}) = -C^{\ell,k}(\mathbf{X}_s) h^{\ell,k}(s).$$

The coefficient $s \mapsto C^{\ell,k}(\mathbf{X}_s) \geq 0$ is continuous, so $E(s) := \exp(\int_0^s C^{\ell,k}(\mathbf{X}_r) dr)$ is C^1 with $E' = C^{\ell,k}(\mathbf{X}_s)E$, and $(h^{\ell,k}E)' = E(h^{\ell,k} + C^{\ell,k}(\mathbf{X}_s)h^{\ell,k}) = 0$. Hence $h^{\ell,k}(s) = h^{\ell,k}(0)\exp(-\int_0^s C^{\ell,k}(\mathbf{X}_r) dr)$. As $C^{\ell,k} \geq 0$, $|h^{\ell,k}(s)|$ is nonincreasing, the sign of $h^{\ell,k}$ is preserved, and $h^{\ell,k}(0) = 0$ implies $h^{\ell,k} \equiv 0$; if this holds for every block, the solution stays in the affine hull, which is therefore invariant.

(iii) Let $\mathbf{x} \in \mathcal{P} \setminus \mathcal{M}_1$ with $x_i^{\ell,k} = 0$. By (i), the i -th coordinate of block (ℓ, k) is

$$b_i^{\ell,k}(\mathbf{x}) = \sum_{i'=1}^{v_k} c_{i'}^{\ell,k}(\mathbf{x}) ((e_{i'})_i - x_i^{\ell,k}) = \sum_{i'} c_{i'}^{\ell,k}(\mathbf{x}) \delta_{ii'} = c_i^{\ell,k}(\mathbf{x}) \geq 0.$$

The polytope $\mathcal{P} = \{\mathbf{x} : \mathbf{1}^\top \mathbf{x}^{\ell,k} = 1, \mathbf{x}^{\ell,k} \geq \mathbf{0} \forall \ell, k\}$ is defined by linear equalities and inequalities, so its (Bouligand) tangent cone at $\mathbf{x} \in \mathcal{P}$ is $T_{\mathcal{P}}(\mathbf{x}) = \{\mathbf{d} : \mathbf{1}^\top \mathbf{d}^{\ell,k} = 0 \forall \ell, k, d_i^{\ell,k} \geq 0 \text{ whenever } x_i^{\ell,k} = 0\}$: if $\mathbf{d}$ lies in the right-hand side, then $\mathbf{x} + h\mathbf{d} \in \mathcal{P}$ for all small $h > 0$ (coordinates with $x_i^{\ell,k} > 0$ stay positive), so $\mathbf{d} \in T_{\mathcal{P}}(\mathbf{x})$; conversely, if $\mathbf{x} + h_n \mathbf{d}_n \in \mathcal{P}$ with $h_n \downarrow 0$ and $\mathbf{d}_n \rightarrow \mathbf{d}$, then $\mathbf{1}^\top \mathbf{d}_n^{\ell,k} = 0$ and $d_{n,i}^{\ell,k} = (x_i^{\ell,k} + h_n d_{n,i}^{\ell,k})/h_n \geq 0$ when $x_i^{\ell,k} = 0$, and both relations pass to the limit. For $\mathbf{x} \in \mathcal{P} \setminus \mathcal{M}_1$, (ii) gives $\mathbf{1}^\top \mathbf{b}^{\ell,k}(\mathbf{x}) = -C^{\ell,k}(\mathbf{x}) \cdot \mathbf{0} = 0$ (here $\mathbf{1}^\top \mathbf{x}^{\ell,k} = 1$), and together with the display above $\mathbf{b}(\mathbf{x}) \in T_{\mathcal{P}}(\mathbf{x})$.

Forward invariance. Let $\mathbf{X}_s, s \in [0, S)$, be a solution in $\mathbb{R}^D \setminus \mathcal{M}_1$ with $\mathbf{X}_0 \in \mathcal{P}$. By (ii), $\mathbf{1}^\top \mathbf{X}_s^{\ell,k} = 1$ for all s and blocks. For the sign constraints, (i) gives for every coordinate

$$\dot{X}_{s,i}^{\ell,k} = b_i^{\ell,k}(\mathbf{X}_s) = c_i^{\ell,k}(\mathbf{X}_s) - C^{\ell,k}(\mathbf{X}_s) X_{s,i}^{\ell,k},$$

so with E from (ii), $(X_{s,i}^{\ell,k} E(s))' = E(s) c_i^{\ell,k}(\mathbf{X}_s) \geq 0$ and $X_{s,i}^{\ell,k} \geq X_{0,i}^{\ell,k}/E(s) \geq 0$. Hence $\mathbf{X}_s \in \mathcal{P}$ for all $s \in [0, S)$. This is the conclusion of Nagumo's theorem [7, 32] for the locally Lipschitz field $\mathbf{b}$ on the open set $\mathcal{O} = \mathbb{R}^D \setminus \mathcal{M}_1$ (it is C^∞ there) and the relatively closed set $\mathcal{P} \cap \mathcal{O}$, on which $\mathbf{b}$ lies in the tangent cone; the computation above makes it explicit. $\square$

B.5 Proof of Proposition 3.5

Proof. Idea. With the product $\prod_k v_k = V$ fixed, the sum $\sum_k v_k$ is smallest when the factors are equal, which is the arithmetic–geometric mean inequality. The remaining one-variable problem is simpler on the logarithmic scale, where $\ln(KV^{1/K}) = \ln K + \ln V/K$ has a derivative whose sign can be read off directly.

By the arithmetic–geometric mean inequality, $\frac{1}{K} \sum_k v_k \geq (\prod_k v_k)^{1/K} = V^{1/K}$, with equality iff all v_k are equal, i.e. iff $v_k = V^{1/K}$ for all k ; multiplying by LK gives $L \sum_k v_k \geq LKV^{1/K}$ with the same equality case. For the second claim (with $V \geq 2$, so that $\ln V > 0$), let $f(K) = KV^{1/K}$ and $g(K) = \ln f(K) = \ln K + \frac{\ln V}{K}$ for real $K > 0$. Then

$$g'(K) = \frac{1}{K} - \frac{\ln V}{K^2} = \frac{K - \ln V}{K^2},$$

which is negative for $K < \ln V$ and positive for $K > \ln V$. So g is strictly decreasing on $(0, \ln V]$ and strictly increasing on $[\ln V, \infty)$, and g , hence $f = e^g$, attains its minimum over $(0, \infty)$ at $K = \ln V$, where $f(\ln V) = \ln V \cdot V^{1/\ln V} = \ln V \cdot e^{\ln V / \ln V} = e \ln V$. $\square$

B.6 Proofs for Section 3.3

The proof of Theorem 3.8 has a local and a global part. Near a vertex, Lemma B.4 estimates the speed M and the target $\bar{\mathbf{p}}$ of Lemma 3.6, and Theorem 3.7 turns these estimates into absorption in finite time, with Lemma B.5 handling the non-smooth distance V_{j*} . Globally, Lemma B.6 shows that the current $q\mathbf{b}$ carries μ_0 to μ_1 , Theorem B.7 decomposes it into curves from noise to data, and Lemma B.8 together with Theorem 3.7 identifies these curves with trajectories that end at a vertex.

Throughout, $\tau := 1 - t$, blocks are indexed by (ℓ, k) , and for a configuration $\mathbf{j}$ we use

$$Z_j(\mathbf{x}) := \pi_j \int_0^1 \frac{p(\mathbf{x} | t, \mathbf{j})}{1 - t} dt, \quad Q_j(\mathbf{x}) := \pi_j \int_0^1 p(\mathbf{x} | t, \mathbf{j}) dt, \quad (43)$$

with $Z_j = Q_j := 0$ if $\pi_j = 0$. Since $w_j = \pi_j p/q$ and q does not depend on t ,

$$\int_0^1 \sum_j \frac{w_j(t, \mathbf{x})}{1 - t} dt = \frac{\sum_j Z_j(\mathbf{x})}{q(\mathbf{x})}, \quad \int_0^1 \sum_j \frac{w_j(t, \mathbf{x})}{1 - t} e_{j_{\ell,k}} dt = \frac{\sum_j Z_j(\mathbf{x}) e_{j_{\ell,k}}}{q(\mathbf{x})},$$

so that, by Eq. (6) and the definition of q , $q = \sum_j Q_j$, $M = \sum_j Z_j / \sum_j Q_j$ and $\bar{\mathbf{p}}^{\ell,k} = \sum_j Z_j \mathbf{e}_{j\ell,k} / \sum_j Z_j$ (whenever $\sum_j Z_j < \infty$, which is shown below for $\mathbf{x} \in \Omega$).

Proof of Lemma 3.6. Idea. None of the three claims needs a new estimate. Configurations with $\pi_j = 0$ contribute nothing to q , $\mathbf{b}$ or Z_j , so the proofs of Lemma B.3 and Proposition 3.3 only need $\mathbf{x}$ to avoid the vertices that carry mass, and this extends them from $\mathbb{R}^D \setminus \mathcal{M}_1$ to Ω . The bounds on M come from comparing Z_j with Q_j , and Eq. (7) is Eq. (5) split into the term with $\mathbf{e}_{j\ell,k}$ and the term with $\mathbf{x}^{\ell,k}$.

Domain. In Lemma B.3(ii)–(iv) and in the proof of Proposition 3.3, the condition $\mathbf{x} \neq \mathbf{e}_j$ is used only for configurations with $\pi_j > 0$: every term with $\pi_j = 0$ vanishes identically. In Lemma B.3(iv) we may therefore take $d := \min_{j:\pi_j>0} \max_{\ell,k} \|\bar{\mathbf{x}}^{\ell,k} - \mathbf{e}_{j\ell,k}\|$, which is positive for $\bar{\mathbf{x}} \in \Omega$ (since $\bar{\mathbf{x}} \neq \mathbf{e}_j$ for every j with $\pi_j > 0$), and obtain the bounds of Lemma B.3(iv) for all j with $\pi_j > 0$. Hence all statements of Lemma B.3 and Proposition 3.3 hold with $\mathbb{R}^D \setminus \mathcal{M}_1$ replaced by Ω , with Eq. (5) as the definition of $\mathbf{b}$ on Ω ; in particular $\mathbf{b} \in C^\infty(\Omega)$, $q > 0$ on $\mathbb{R}^D$, and $Z_j(\mathbf{x}) < \infty$ for $\mathbf{x} \in \Omega$ and all j by Lemma B.3(iii) (divide by a_1). Since $\frac{1}{1-t} \leq \frac{1}{a_1}(1 + \frac{a_1}{1-t})$, the integrands of Z_j are dominated near each $\bar{\mathbf{x}} \in \Omega$ by G_β/a_1 together with their $\mathbf{x}$ -derivatives, so the argument of the smoothness part of the proof of Proposition 3.3 shows that Z_j , and hence M and $\bar{\mathbf{p}}^{\ell,k}$, are C^∞ on Ω .

Speed. For $\mathbf{x} \in \Omega$, $\sum_j Z_j(\mathbf{x}) < \infty$ and $q(\mathbf{x}) \in (0, \infty)$, so $M(\mathbf{x}) = \sum_j Z_j(\mathbf{x})/q(\mathbf{x}) < \infty$. Since $\frac{1}{1-t} \geq 1$ on $[0, 1)$, $Z_j \geq Q_j$ for every j , so $\sum_j Z_j \geq \sum_j Q_j = q > 0$ and $M(\mathbf{x}) \geq 1$. The weights $Z_j(\mathbf{x})/\sum_{j'} Z_{j'}(\mathbf{x})$ are nonnegative and sum to one, so $\bar{\mathbf{p}}^{\ell,k}(\mathbf{x}) = \sum_j \frac{Z_j}{\sum_{j'} Z_{j'}} \mathbf{e}_{j\ell,k}$ is a convex combination of vertices of Δ^{v_k-1} .

Formula. Both integrals below converge absolutely: the second equals $M(\mathbf{x}) < \infty$, and the integrand of the first has norm at most that of the second ($\|\mathbf{e}_{j\ell,k}\| = 1$). Splitting Eq. (5) accordingly,

$$\mathbf{b}^{\ell,k}(\mathbf{x}) = a_k \int_0^1 \sum_j \frac{w_j(t, \mathbf{x})}{1-t} \mathbf{e}_{j\ell,k} dt - a_k \left(\int_0^1 \sum_j \frac{w_j(t, \mathbf{x})}{1-t} dt \right) \mathbf{x}^{\ell,k} = a_k M(\mathbf{x}) \bar{\mathbf{p}}^{\ell,k}(\mathbf{x}) - a_k M(\mathbf{x}) \mathbf{x}^{\ell,k},$$

where the last step uses the definitions Eq. (6). $\square$

Notation near a vertex. Fix $\mathbf{e}_{j^*} \in \mathcal{M}_1^+$, write $\mathbf{e}_*^{\ell,k} := \mathbf{e}_{j^*\ell,k}$, $\boldsymbol{\delta}^{\ell,k} := \mathbf{x}^{\ell,k} - \mathbf{e}_*^{\ell,k}$, $V := V_{j^*}(\mathbf{x}) = \max_{\ell,k} \|\boldsymbol{\delta}^{\ell,k}\|^{1/a_k}$, $Z_* := Z_{j^*}$, $Q_* := Q_{j^*}$, $Z_c := \sum_{j \neq j^*} Z_j$, $Q_c := \sum_{j \neq j^*} Q_j$, $a_{\min} = \min_k a_k$, $a_{\max} = \max_k a_k$, $\sigma_{\max} = \max_k \sigma_k$, and $\kappa_0 := \prod_{\ell,k} (2\pi\sigma_k^2)^{-v_k/2}$. By definition of V , raising $\|\boldsymbol{\delta}^{\ell,k}\|^{1/a_k} \leq V$ to the power $a_k > 0$,

$$\|\boldsymbol{\delta}^{\ell,k}\| \leq V^{a_k} \quad \text{for all } (\ell, k), \quad \text{with equality for at least one block.} \quad (44)$$

Since $\alpha_k = \tau^{a_k}$ and $\prod_{\ell,k} \alpha_k^{-v_k} = \tau^{-N}$ (Eq. (39)), Eq. (4) reads

$$p(\mathbf{x} \mid t, j^*) = \kappa_0 \tau^{-N} \exp\left(-\sum_{\ell,k} \frac{\|\boldsymbol{\delta}^{\ell,k} + \alpha_k \mathbf{e}_*^{\ell,k}\|^2}{2\alpha_k^2 \sigma_k^2}\right), \quad (45)$$

because $\mathbf{x}^{\ell,k} - (1 - \alpha_k) \mathbf{e}_*^{\ell,k} = \boldsymbol{\delta}^{\ell,k} + \alpha_k \mathbf{e}_*^{\ell,k}$. If $v_k \geq 2$ for every k (as in Theorem 3.8), then, since $L \geq 1$ and all terms of $N = L \sum_k a_k v_k$ are positive, $N \geq L v_k a_k \geq 2a_k$ for every k , hence

$$\gamma := N - a_{\max} \geq a_{\max} > 0. \quad (46)$$

Without the assumption $v_k \geq 2$, still $\gamma > 0$ unless $L = K = v_1 = 1$: if $L \geq 2$ or $K \geq 2$ or $v_1 \geq 2$, then N contains, besides one term $a_{\max}$, at least one further positive term. In the remaining case $D = 1$, $\mathcal{J}$ has a single element, so $Z_c \equiv 0$ and $\bar{\mathbf{p}}^{\ell,k} \equiv \mathbf{e}_*^{\ell,k}$.

Lemma B.4 (Bounds near a vertex). *There are constants $\varepsilon_0 \in (0, 1)$ and $C_0 \geq 1$ (depending on π_{j^*} , (a_k, σ_k, v_k) , L) such that for all $\mathbf{x}$ with $0 < V_{j^*}(\mathbf{x}) < \varepsilon_0$:*

- (a) $Z_c(\mathbf{x}) \leq C_0$ and $Q_c(\mathbf{x}) \leq C_0$;
- (b) $C_0^{-1} V^{-N} \leq Z_*(\mathbf{x}) \leq C_0 V^{-N}$;

- (c) $Q_*(\mathbf{x}) \geq C_0^{-1}V^{1-N}$ if $N > 1$; and $Q_*(\mathbf{x}) \leq C_0V^{1-N}$ if $N > 1$, $Q_*(\mathbf{x}) \leq C_0(1 + \log(1/V))$ if $N \leq 1$;
- (d) $\|\bar{\mathbf{p}}^{\ell,k}(\mathbf{x}) - \mathbf{e}_*^{\ell,k}\| \leq C_0V^N$ for every block;
- (e) $M(\mathbf{x}) \geq C_0^{-1}V^{-1}$ and $M(\mathbf{x}) \leq C_0V^{-1}$ if $N > 1$; $M(\mathbf{x}) \geq C_0^{-1}V^{-N}/(1 + \log(1/V))$ if $N \leq 1$.

Proof. Idea. All five bounds concern the four integrals Z_*, Q_*, Z_c, Q_c , since M and $\bar{\mathbf{p}}$ are ratios of them. A competing configuration differs from $\mathbf{j}^*$ in some block by a distance close to $\sqrt{2}$, so its likelihood at $\mathbf{x}$ is exponentially small as $\tau \rightarrow 0$ and its integral stays bounded, which is (a). For $\mathbf{j}^*$ itself, V is the natural time scale: for $\tau \geq V$ every deviation $\|\delta^{\ell,k}\|$ is at most α_k and $p(\mathbf{x} | t, \mathbf{j}^*)$ is comparable to τ^{-N} , while for $\tau \ll V$ the block attaining V lies many standard deviations away and the likelihood is negligible; the substitution $\tau = V\xi$ turns this into (b) and (c). Parts (d) and (e) then follow by taking ratios.

Choose

$$\varepsilon_0 := \min \left\{ \min_k 4^{-1/a_k}, 2^{-1/N}, 2^{-1/(N-1)} \text{ (the last only if } N > 1) \right\} \in (0, 1).$$

Then $V < \varepsilon_0$ implies $\|\delta^{\ell,k}\| \leq V^{a_k} \leq 1/4$ for all blocks by Eq. (44), $V^{-N} \geq 2$, and $V^{1-N} \geq 2$ if $N > 1$. In particular $\mathbf{x} \in \Omega$ (indeed $\mathbf{x} \notin \mathcal{M}_1$): $\mathbf{x} \neq \mathbf{e}_{j^*}$ as $V > 0$, and $\mathbf{x} \neq \mathbf{e}_j$ for $j \neq j^*$ because some block has $\|\mathbf{e}_{j_{\ell,k}} - \mathbf{e}_*^{\ell,k}\| = \sqrt{2} > 1/4$. So M and $\bar{\mathbf{p}}$ are defined at $\mathbf{x}$. In integrals over t we substitute $\tau = 1 - t$, $dt = -d\tau$, so $\int_0^1 f(t) dt = \int_0^1 f(1 - \tau) d\tau$. Constants $C_1, C_2, \dots$ depend only on the data listed in the statement.

(a) Let $\mathbf{j} \neq \mathbf{j}^*$ with $\pi_j > 0$ (the other terms vanish), and pick a block (ℓ_0, k_0) with $j_{\ell_0, k_0} \neq j_{\ell_0, k_0}^*$. Write $\mathbf{e} := \mathbf{e}_{j_{\ell_0, k_0}}$, $\alpha := \alpha_{k_0}$. Two distinct vertices of a simplex are at distance $\sqrt{2}$, so by the triangle inequality, $\|\delta^{\ell_0, k_0}\| \leq 1/4$ and $\|\mathbf{e}\| = 1$,

$$\|\mathbf{x}^{\ell_0, k_0} - (1 - \alpha)\mathbf{e}\| = \|(\mathbf{e}_*^{\ell_0, k_0} - \mathbf{e}) + \delta^{\ell_0, k_0} + \alpha\mathbf{e}\| \geq \sqrt{2} - \frac{1}{4} - \alpha.$$

Let $\tau_1 := \min_k 2^{-1/a_k}$. For $\tau \leq \tau_1$ all $\alpha_k = \tau^{a_k} \leq \tau_1^{a_k} \leq 1/2$, so the norm is at least $\sqrt{2} - 3/4 > 1/2$, and $\alpha = \tau^{a_{k_0}} \leq \tau^{a_{\min}}$ (as $\tau \leq 1$). The factor of block (ℓ_0, k_0) in Eq. (4) is therefore at most $\exp(-\frac{1}{8\alpha^2\sigma_{k_0}^2}) \leq \exp(-\frac{1}{8\sigma_{\max}^2\tau^{2a_{\min}}})$. Bounding the other exponential factors of Eq. (4) by one and using Eq. (39),

$$\frac{p(\mathbf{x} | t, \mathbf{j})}{1 - t} \leq \kappa_0 \tau^{-N-1} \exp\left(-\frac{1}{8\sigma_{\max}^2\tau^{2a_{\min}}}\right) \quad (\tau \leq \tau_1).$$

The right-hand side is integrable on $(0, \tau_1]$: substituting $y = 1/\tau$, $d\tau = y^{-2}dy$,

$$\int_0^{\tau_1} \tau^{-N-1} e^{-\tau^{-2a_{\min}}/(8\sigma_{\max}^2)} d\tau = \int_{1/\tau_1}^{\infty} y^{N-1} e^{-y^{2a_{\min}}/(8\sigma_{\max}^2)} dy < \infty.$$

For $\tau \in [\tau_1, 1]$, $\alpha_k \geq \tau_1^{a_k}$, so $p(\mathbf{x} | t, \mathbf{j}) \leq \prod_{\ell,k} (2\pi\tau_1^{2a_k}\sigma_k^2)^{-v_k/2}$ and $1/(1-t) \leq 1/\tau_1$. Both bounds are independent of $\mathbf{x}$, so (with $\pi_j \leq 1$) $Z_j(\mathbf{x}) \leq C_1$, and summing over the finitely many $\mathbf{j}$ gives $Z_c \leq |\mathcal{J}|C_1$. Since $p \leq p/(1-t)$, also $Q_c \leq Z_c$.

(b), lower bound. For $\tau \in [V, 1]$ and every block, $\alpha_k = \tau^{a_k} \geq V^{a_k} \geq \|\delta^{\ell,k}\|$ by Eq. (44), so $\|\delta^{\ell,k} + \alpha_k \mathbf{e}_*^{\ell,k}\| \leq \|\delta^{\ell,k}\| + \alpha_k \leq 2\alpha_k$ and the exponent in Eq. (45) is at least $-\sum_{\ell,k} 4\alpha_k^2/(2\alpha_k^2\sigma_k^2) = -\sum_{\ell,k} 2/\sigma_k^2 =: -c_0$. Hence $p(\mathbf{x} | t, \mathbf{j}^*) \geq \kappa_0 e^{-c_0} \tau^{-N}$ on $[V, 1]$ and, dropping the (nonnegative) contribution of $\tau < V$,

$$Z_* \geq \pi_{j^*} \kappa_0 e^{-c_0} \int_V^1 \tau^{-N-1} d\tau = \pi_{j^*} \kappa_0 e^{-c_0} \frac{V^{-N} - 1}{N} \geq \frac{\pi_{j^*} \kappa_0 e^{-c_0}}{2N} V^{-N},$$

using $V^{-N} - 1 \geq \frac{1}{2}V^{-N}$, which is $V^{-N} \geq 2$.

(b), upper bound. Let (ℓ^*, k^*) be a block with equality in Eq. (44), so $\|\delta^{\ell^*, k^*}\| = V^{a_{k^*}} > 0$, and set $\tau_2 := 2^{-1/a_{\min}}V < V$. For $\tau \leq \tau_2$, since $a_{k^*}/a_{\min} \geq 1$,

$$\alpha_{k^*} = \tau^{a_{k^*}} \leq 2^{-a_{k^*}/a_{\min}} V^{a_{k^*}} \leq \frac{1}{2} V^{a_{k^*}} = \frac{1}{2} \|\delta^{\ell^*, k^*}\|,$$

so $\|\delta^{\ell^*, k^*} + \alpha_{k^*} \mathbf{e}_{\star}^{\ell^*, k^*}\| \geq \|\delta^{\ell^*, k^*}\| - \alpha_{k^*} \geq \frac{1}{2} V^{a_{k^*}}$. The factor of this block in Eq. (45) is then at most $\exp(-V^{2a_{k^*}}/(8\tau^{2a_{k^*}}\sigma_{k^*}^2))$, and bounding all other factors by one,

$$p(\mathbf{x} | t, \mathbf{j}^*) \leq \kappa_0 \tau^{-N} \exp\left(-\frac{(V/\tau)^{2a_{k^*}}}{8\sigma_{\max}^2}\right) \quad (\tau \leq \tau_2).$$

Substituting $\tau = V\xi$, $d\tau = Vd\xi$, $\tau^{-N-1}d\tau = V^{-N}\xi^{-N-1}d\xi$,

$$\int_0^{\tau_2} \frac{p}{\tau} d\tau \leq \kappa_0 V^{-N} \int_0^{2^{-1/a_{\min}}} \xi^{-N-1} e^{-\xi^{-2a_{k^*}}/(8\sigma_{\max}^2)} d\xi \leq C_2 V^{-N},$$

where C_2 is κ_0 times the maximum of the integrals over the finitely many values $a_{k^*} \in \{a_1, \dots, a_K\}$; each integral is finite, as the substitution $y = 1/\xi$ turns it into $\int_{2^{1/a_{\min}}}^{\infty} y^{N-1} e^{-y^{2a_{k^*}}/(8\sigma_{\max}^2)} dy$. On $[\tau_2, 1]$, bounding the exponential in Eq. (45) by one, $p \leq \kappa_0 \tau^{-N}$ and $\int_{\tau_2}^1 \tau^{-N-1} d\tau \leq \tau_2^{-N}/N = 2^{N/a_{\min}} V^{-N}/N$. With $\pi_{\mathbf{j}^*} \leq 1$, $Z_{\star} \leq (C_2 + \kappa_0 2^{N/a_{\min}}/N) V^{-N}$.

(c) The same two estimates without the factor $1/\tau$ give the following. Lower bound for $N > 1$: on $[V, 1]$,

$$Q_{\star} \geq \pi_{\mathbf{j}^*} \kappa_0 e^{-c_0} \int_V^1 \tau^{-N} d\tau = \pi_{\mathbf{j}^*} \kappa_0 e^{-c_0} \frac{V^{1-N} - 1}{N-1} \geq \frac{\pi_{\mathbf{j}^*} \kappa_0 e^{-c_0}}{2(N-1)} V^{1-N},$$

using $V^{1-N} \geq 2$. Upper bound: substituting $\tau = V\xi$ on $(0, \tau_2]$ ($\tau^{-N} d\tau = V^{1-N} \xi^{-N} d\xi$),

$$Q_{\star} \leq \kappa_0 V^{1-N} \int_0^{2^{-1/a_{\min}}} \xi^{-N} e^{-\xi^{-2a_{k^*}}/(8\sigma_{\max}^2)} d\xi + \kappa_0 \int_{\tau_2}^1 \tau^{-N} d\tau.$$

The first integral is finite (as in (b)) and bounded by a constant C_3 . If $N > 1$, the last integral is at most $\tau_2^{1-N}/(N-1) = 2^{(N-1)/a_{\min}} V^{1-N}/(N-1)$, so $Q_{\star} \leq C_4 V^{1-N}$. If $N \leq 1$, then $\tau^{-N} \leq \tau^{-1}$ on $(0, 1]$ and the last integral is at most $\int_{\tau_2}^1 \tau^{-1} d\tau = \log(1/\tau_2) = \log(1/V) + \frac{\log 2}{a_{\min}}$; the first term is at most $\kappa_0 C_3 V^{1-N} \leq \kappa_0 C_3$ because $V < 1$ and $1-N \geq 0$. Hence $Q_{\star} \leq C_5(1 + \log(1/V))$ if $N \leq 1$.

Let $C'_1 \geq 1$ be the largest of the constants obtained so far in (a)–(c) (for upper bounds) and of their reciprocals (for lower bounds). Then (a)–(c) hold with C'_1 in place of C_0 .

(d) By the formula for $\bar{\mathbf{p}}^{\ell, k}$ after Eq. (43) and $\sum_{\mathbf{j}} Z_{\mathbf{j}} \mathbf{e}_{\star}^{\ell, k} / \sum_{\mathbf{j}} Z_{\mathbf{j}} = \mathbf{e}_{\star}^{\ell, k}$,

$$\bar{\mathbf{p}}^{\ell, k} - \mathbf{e}_{\star}^{\ell, k} = \frac{\sum_{\mathbf{j}} Z_{\mathbf{j}} (\mathbf{e}_{j_{\ell, k}} - \mathbf{e}_{\star}^{\ell, k})}{\sum_{\mathbf{j}} Z_{\mathbf{j}}} = \frac{\sum_{\mathbf{j} \neq \mathbf{j}^*} Z_{\mathbf{j}} (\mathbf{e}_{j_{\ell, k}} - \mathbf{e}_{\star}^{\ell, k})}{Z_{\star} + Z_c},$$

and $\|\mathbf{e}_{j_{\ell, k}} - \mathbf{e}_{\star}^{\ell, k}\| \leq \sqrt{2}$. By the triangle inequality, (a) and (b), $\|\bar{\mathbf{p}}^{\ell, k} - \mathbf{e}_{\star}^{\ell, k}\| \leq \sqrt{2} Z_c / (Z_{\star} + Z_c) \leq \sqrt{2} Z_c / Z_{\star} \leq \sqrt{2} C_1'^2 V^N$.

(e) $M = (Z_{\star} + Z_c) / (Q_{\star} + Q_c)$. Lower bound: $M \geq Z_{\star} / (Q_{\star} + Q_c)$. For $N > 1$, $V < 1$ gives $V^{1-N} > 1$, so by (a) and (c), $Q_{\star} + Q_c \leq C'_1 V^{1-N} + C'_1 \leq 2C'_1 V^{1-N}$, and with (b)

$$M \geq \frac{C_1'^{-1} V^{-N}}{2C_1' V^{1-N}} = \frac{V^{-1}}{2C_1'^2}.$$

For $N \leq 1$, $Q_{\star} + Q_c \leq C'_1(1 + \log(1/V)) + C'_1 \leq 2C'_1(1 + \log(1/V))$ (as $\log(1/V) > 0$), so $M \geq V^{-N} / (2C_1'^2(1 + \log(1/V)))$. Upper bound for $N > 1$: using $V^{-N} \geq 1$,

$$M \leq \frac{Z_{\star} + Z_c}{Q_{\star}} \leq \frac{C_1' V^{-N} + C_1'}{C_1'^{-1} V^{1-N}} \leq \frac{2C_1' V^{-N}}{C_1'^{-1} V^{1-N}} = 2C_1'^2 V^{-1}.$$

Set $C_0 := 2C_1'^2 \geq C'_1$. All upper bounds in (a)–(e) are of the form (constant at most C_0) $\times$ (function of V) and all lower bounds of the form (constant at least C_0^{-1}) $\times$ (function of V), so (a)–(e) hold with C_0 . $\square$

Lemma B.5 (Dini derivatives). *Let $f : [s_1, s_2] \rightarrow \mathbb{R}$ be continuous with upper right Dini derivative $D^+f(s) := \limsup_{h \downarrow 0} (f(s+h) - f(s))/h \leq -\kappa$ for all $s \in [s_1, s_2]$. Then $f(s) \leq f(s_1) - \kappa(s - s_1)$ on $[s_1, s_2]$. If $f = \max_{i \in I} f_i$ for finitely many functions f_i that are C^1 near s , then $D^+f(s) \leq \max_{i \in I(s)} f'_i(s)$ and $D_+f(s) := \liminf_{h \downarrow 0} (f(s+h) - f(s))/h \geq \max_{i \in I(s)} f'_i(s)$, where $I(s) = \{i : f_i(s) = f(s)\}$.*

Proof. Idea. The distance V_{j^*} of Theorem 3.7 is a maximum of finitely many smooth functions and is not differentiable where the maximum switches, so its monotonicity has to be argued with one-sided derivatives. The first claim is the mean value inequality for the upper right Dini derivative, proved by looking at the largest point up to which a slightly weaker linear bound holds. The second claim reduces to the observation that near s only the indices active at s matter, and a maximum of finitely many C^1 functions has a right derivative.

Here $\kappa \in \mathbb{R}$ is arbitrary. Fix $\kappa' < \kappa$ and let $S = \{s \in [s_1, s_2] : f(s) \leq f(s_1) - \kappa'(s - s_1)\}$; it is closed because f is continuous, and it contains s_1 . Let $s^* := \max S$ and suppose $s^* < s_2$. Since $D^+f(s^*) \leq -\kappa < -\kappa'$, there is $h_0 > 0$ with $(f(s^* + h) - f(s^*))/h < -\kappa'$ for all $h \in (0, h_0)$, hence

$$f(s^* + h) < f(s^*) - \kappa'h \leq f(s_1) - \kappa'(s^* - s_1) - \kappa'h = f(s_1) - \kappa'(s^* + h - s_1),$$

so $s^* + h \in S$ for small $h > 0$, contradicting the maximality of s^* . Hence $s_2 \in S$. Applying this on $[s_1, s]$ for each $s \in (s_1, s_2]$ gives $f(s) \leq f(s_1) - \kappa'(s - s_1)$, and letting $\kappa' \uparrow \kappa$ gives the claim.

For the maximum: f is continuous near s . An index with $f_i(s) < f(s)$ satisfies $f_i < f$ on a neighbourhood of s by continuity, so near s , $f = \max_{i \in I(s)} f_i$. For $i \in I(s)$, $f_i(s+h) = f(s) + hf'_i(s) + r_i(h)$ with $r_i(h) = o(h)$, so for small $h > 0$

$$\left| \frac{f(s+h) - f(s)}{h} - \max_{i \in I(s)} f'_i(s) \right| \leq \max_{i \in I(s)} \frac{|r_i(h)|}{h} \rightarrow 0.$$

Thus the right derivative of f at s exists and equals $\max_{i \in I(s)} f'_i(s)$, which gives both inequalities. $\square$

Proof of Theorem 3.7. Idea. We need a quantity that decreases along trajectories near e_{j^*} and reaches zero in finite time, and the distance $V = V_{j^*}$ serves as this Lyapunov function. The exponents $1/a_k$ in its definition are matched to Eq. (7): since $\mathbf{b}^{\ell,k} = a_k M(\bar{\mathbf{p}}^{\ell,k} - \mathbf{x}^{\ell,k})$, each $\|\delta^{\ell,k}\|^{1/a_k}$ shrinks at the same relative rate M , up to the error $\bar{\mathbf{p}}^{\ell,k} - \mathbf{e}_*^{\ell,k}$, which is of order V^N by Lemma B.4(d). Step 1 turns this into the differential inequality Eq. (48); Step 2 shows that trajectories stay near the vertex and, because M blows up fast enough as $V \rightarrow 0$ (Lemma B.4(e)), reach it in finite time; Step 3 uses the two-sided bound $M \asymp V^{-1}$ for $N > 1$ to obtain the rates.

Let ε_0, C_0 be as in Lemma B.4. If $\gamma > 0$, choose $\varepsilon \leq \varepsilon_0$ so small that $C_0 \varepsilon^\gamma \leq \frac{1}{2}$, with γ from Eq. (46); in the degenerate case $\gamma = 0$ (then $D = 1$ and $\bar{\mathbf{p}} \equiv \mathbf{e}_*$, see after Eq. (46)) take $\varepsilon := \varepsilon_0$. Let $W := \{\mathbf{x} : 0 < V_{j^*}(\mathbf{x}) < \varepsilon\}$, an open set since V_{j^*} is continuous. Then $W \subset \Omega$: $\mathbf{x} \in W$ is not e_{j^*} (as $V_{j^*}(\mathbf{x}) > 0$), and every other vertex $e_j \in \mathcal{M}_1^+$ has a block with $\|e_{j\ell,k} - \mathbf{e}_*^{\ell,k}\| = \sqrt{2} > \varepsilon^{a_k}$, whereas $\|\mathbf{x}^{\ell,k} - \mathbf{e}_*^{\ell,k}\| < \varepsilon^{a_k}$ on W by Eq. (44); so $W \cap \mathcal{M}_1^+ = \emptyset$. The same argument shows

$$\{V_{j^*} < \varepsilon\} \cap \mathcal{M}_1^+ = \{e_{j^*}\}. \quad (47)$$

By Lemma 3.6, $\mathbf{b}$ is C^∞ , hence locally Lipschitz, on Ω , so for every $\mathbf{x}_0 \in \Omega$ there is a unique maximal solution of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$, $\mathbf{X}_0 = \mathbf{x}_0$, in Ω ; we write $[0, \tau(\mathbf{x}_0))$ for its interval of existence, $\tau(\mathbf{x}_0) \in (0, \infty]$.

Step 1: a differential inequality. Let $\mathbf{X}_s$ be a solution in W on an interval, and $V(s) := V_{j^*}(\mathbf{X}_s)$. For a block with $\delta^{\ell,k}(s) \neq \mathbf{0}$, the map $s \mapsto \|\delta^{\ell,k}(s)\|$ is C^1 near s , and since $\delta^{\ell,k} = \mathbf{X}_s^{\ell,k} = \mathbf{b}^{\ell,k}(\mathbf{X}_s)$, Eq. (7) gives

$$\frac{d}{ds} \|\delta^{\ell,k}\| = \left\langle \frac{\delta^{\ell,k}}{\|\delta^{\ell,k}\|}, \mathbf{b}^{\ell,k}(\mathbf{X}_s) \right\rangle = a_k M \left(-\|\delta^{\ell,k}\| + \left\langle \frac{\delta^{\ell,k}}{\|\delta^{\ell,k}\|}, \bar{\mathbf{p}}^{\ell,k} - \mathbf{e}_*^{\ell,k} \right\rangle \right),$$

because $\bar{\mathbf{p}}^{\ell,k} - \mathbf{x}^{\ell,k} = (\bar{\mathbf{p}}^{\ell,k} - \mathbf{e}_*^{\ell,k}) - \delta^{\ell,k}$ and $\langle \delta^{\ell,k}, \delta^{\ell,k} \rangle / \|\delta^{\ell,k}\| = \|\delta^{\ell,k}\|$. Let $f_{\ell,k} := \|\delta^{\ell,k}\|^{1/a_k}$. By the chain rule, the Cauchy–Schwarz inequality $|\langle \delta^{\ell,k} / \|\delta^{\ell,k}\|, \bar{\mathbf{p}}^{\ell,k} - \mathbf{e}_*^{\ell,k} \rangle| \leq \|\bar{\mathbf{p}}^{\ell,k} - \mathbf{e}_*^{\ell,k}\|$ and Lemma B.4(d) (applicable since $0 < V < \varepsilon \leq \varepsilon_0$),

$$f'_{\ell,k} = \frac{1}{a_k} \|\delta^{\ell,k}\|^{\frac{1}{a_k}-1} \frac{d}{ds} \|\delta^{\ell,k}\| \in M \left[-f_{\ell,k} - \|\delta^{\ell,k}\|^{\frac{1}{a_k}-1} C_0 V^N, -f_{\ell,k} + \|\delta^{\ell,k}\|^{\frac{1}{a_k}-1} C_0 V^N \right].$$

Fix s and let (ℓ, k) be an active block, i.e. $f_{\ell, k}(s) = V(s) > 0$. Then $\delta^{\ell, k}(s) \neq \mathbf{0}$, so $f_{\ell, k}$ is C^1 near s ($\mathbf{X}$ is C^1 , the norm is smooth away from $\mathbf{0}$, and $r \mapsto r^{1/a_k}$ is smooth on $(0, \infty)$), and $\|\delta^{\ell, k}\|^{1/a_k - 1} = V^{1-a_k}$. As $N - a_k \geq N - a_{\max} = \gamma$ and $V < 1$,

$$V^{1-a_k} C_0 V^N = C_0 V \cdot V^{N-a_k} \leq C_0 V \cdot V^\gamma \leq C_0 V \varepsilon^\gamma \leq \frac{1}{2} V$$

(in the degenerate case the error term is 0 because $\bar{\mathbf{p}} = \mathbf{e}_\star$). Hence $f'_{\ell, k}(s) \in M[-\frac{3}{2}V, -\frac{1}{2}V]$ for every active block. Inactive blocks have $f_{\ell, k}(s) < V(s)$, and near s , V is the maximum of the active $f_{\ell, k}$ (proof of Lemma B.5), each C^1 near s . Lemma B.5 applied to this maximum gives

$$D^+V(s) \leq -\frac{1}{2}M(\mathbf{X}_s)V(s), \quad D_+V(s) \geq -\frac{3}{2}M(\mathbf{X}_s)V(s). \quad (48)$$

Step 2: invariance and absorption. Let $\mathbf{x}_0 \in W$, let $[0, s^*)$ with $s^* = \tau(\mathbf{x}_0)$ be the maximal interval of existence of the solution in Ω with $\mathbf{X}_0 = \mathbf{x}_0$, and $s_1 := \sup\{s < s^* : \mathbf{X}_r \in W \ \forall r \leq s\}$; $s_1 > 0$ because W is open and $\mathbf{X}$ continuous. On $[0, s_1]$ the solution lies in W and Eq. (48) holds. On a compact subinterval $[s', s''] \subset [0, s_1]$ the function $s \mapsto \frac{1}{2}M(\mathbf{X}_s)V(s)$ is continuous (M is continuous on Ω by Lemma 3.6) and positive, so it has a positive minimum κ , and Lemma B.5 gives $V(s'') \leq V(s') - \kappa(s'' - s') < V(s')$. So V is strictly decreasing on $[0, s_1]$, whence $V(s) \leq V(0) < \varepsilon$. Suppose $s_1 < s^*$. By continuity $V(s_1) = \lim_{s \uparrow s_1} V(s) \leq V(0) < \varepsilon$, and $V(s_1) > 0$ because $\mathbf{X}_{s_1} \in \Omega$ and $V_{j^*} = 0$ only at $\mathbf{e}_{j^*} \notin \Omega$. So $\mathbf{X}_{s_1} \in W$ and, W being open, $\mathbf{X}_r \in W$ for all $r \leq s_1 + h$ for some $h > 0$, contradicting the definition of s_1 . So $s_1 = s^*$, $\mathbf{X}_s \in W$ for all $s < s^*$, and V is strictly decreasing on $[0, s^*)$.

Finite time. By Lemma B.4(e), $M \geq m(V)$ with $m(V) = C_0^{-1}V^{-1}$ if $N > 1$ and $m(V) = C_0^{-1}V^{-N}/(1 + \log(1/V))$ if $N \leq 1$. Let $G(v) := \int_0^v du/(\frac{1}{2}um(u))$ for $v \in (0, \varepsilon)$. The integrand is $2C_0$ (if $N > 1$) or $2C_0u^{N-1}(1 + \log(1/u))$ (if $N \leq 1$), both continuous and positive on $(0, \varepsilon)$ and integrable at 0 as $N > 0$; so G is finite, $G \geq 0$, and G is C^1 and increasing on $(0, \varepsilon)$ with $G'(v) = 2/(vm(v)) > 0$. For $s < s^*$ and $h > 0$ small, the mean value theorem gives $G(V(s+h)) - G(V(s)) = G'(\xi_h)(V(s+h) - V(s))$ with ξ_h between $V(s)$ and $V(s+h)$; as $h \downarrow 0$, $G'(\xi_h) \rightarrow G'(V(s)) > 0$ by continuity of V and G' . Hence, by Eq. (48) and $M \geq m(V)$,

$$D^+(G \circ V)(s) = G'(V(s)) D^+V(s) \leq -\frac{2}{Vm(V)} \cdot \frac{1}{2}M(\mathbf{X}_s)V = -\frac{M(\mathbf{X}_s)}{m(V(s))} \leq -1.$$

Lemma B.5 on $[0, s]$ gives $G(V(s)) \leq G(V(0)) - s$. Since $G \geq 0$, $s \leq G(V(0))$ for every $s < s^*$, so $s^* \leq G(V(0)) < \infty$.

Limit. For $\eta \in (0, V(0))$ let $K_\eta := \{\mathbf{x} : \eta \leq V_{j^*}(\mathbf{x}) \leq V(0)\}$. It is closed (V_{j^*} is continuous) and bounded ($\|\mathbf{x}^{\ell, k}\| \leq 1 + V(0)^{a_k}$ on it by Eq. (44)), hence compact, and it is contained in Ω (by Eq. (47), as $V_{j^*} \geq \eta > 0$ on it and $V(0) < \varepsilon$). A maximal solution of an ODE with locally Lipschitz right-hand side on an open set Ω with finite existence time s^* eventually leaves every compact subset of Ω : there is $s_\eta < s^*$ with $\mathbf{X}_s \notin K_\eta$ for $s \in (s_\eta, s^*)$. Since $V(s) \leq V(0)$, this means $V(s) < \eta$ for $s \in (s_\eta, s^*)$. Hence $V(s) \rightarrow 0$, and by Eq. (44) $\|\mathbf{X}_s^{\ell, k} - \mathbf{e}_\star^{\ell, k}\| \leq V(s)^{a_k} \rightarrow 0$, i.e. $\mathbf{X}_s \rightarrow \mathbf{e}_{j^*}$ as $s \uparrow s^* = \tau(\mathbf{x}_0)$. This proves (i).

Step 3: rates for $N > 1$. By Lemma B.4(e), $C_0^{-1}V^{-1} \leq M \leq C_0V^{-1}$, so Eq. (48) gives

$$D^+V \leq -\frac{1}{2}C_0^{-1}V^{-1}V = -\frac{1}{2C_0} =: -c, \quad D_+V \geq -\frac{3}{2}C_0V^{-1}V = -\frac{3}{2}C_0 =: -C,$$

and $0 < c \leq C$ as $C_0 \geq 1$. Lemma B.5 applied to V with $\kappa = c$ gives $V(s) \leq V(0) - cs$. Since $D^+(-V) = -D_+V \leq C$, Lemma B.5 applied to $-V$ with $\kappa = -C$ gives $-V(s) \leq -V(0) + Cs$, i.e. $V(s) \geq V(0) - Cs$. Letting $s \uparrow \tau(\mathbf{x}_0)$, where $V \rightarrow 0$: from $0 \leq V(s) \leq V(0) - cs$ we get $s \leq V(0)/c$ for all $s < \tau(\mathbf{x}_0)$, so $\tau(\mathbf{x}_0) \leq V(0)/c$; from $V(s) \geq V(0) - Cs$ we get $0 \geq V(0) - C\tau(\mathbf{x}_0)$, so $\tau(\mathbf{x}_0) \geq V(0)/C$. For fixed $s < \tau(\mathbf{x}_0)$, Lemma B.5 applied to $-V$ on $[s, s']$ gives $V(s') \geq V(s) - C(s' - s)$; letting $s' \uparrow \tau(\mathbf{x}_0)$ yields $V(s) \leq C(\tau(\mathbf{x}_0) - s)$, and Eq. (44) gives $\|\mathbf{X}_s^{\ell, k} - \mathbf{e}_\star^{\ell, k}\| \leq V(s)^{a_k} \leq (C(\tau(\mathbf{x}_0) - s))^{a_k}$. $\square$

Lemma B.6 (Continuity equation). *The current $\mathbf{j} := q\mathbf{b}$ lies in $L^1(\mathbb{R}^D; \mathbb{R}^D)$ and $\nabla \cdot \mathbf{j} = \mu_0 - \mu_1$ in the sense of distributions: $\int \nabla \phi \cdot \mathbf{j} d\mathbf{x} = \int \phi d\mu_1 - \int \phi d\mu_0$ for all $\phi \in C_c^\infty(\mathbb{R}^D)$.*

Proof. Idea. The identity is a statement about the interpolant averaged over time, and it is easiest to prove it path by path and average afterwards. Along a single path the fundamental theorem of calculus gives $\phi(\mathbf{x}_1) - \phi(\mathbf{x}_0) = \int_0^1 \nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t dt$; taking expectations and replacing $\dot{\mathbf{I}}_t$ by its conditional expectation $\mathbf{b}(\mathbf{I}_t)$ (tower property) produces $\int \nabla \phi \cdot \mathbf{q} \mathbf{b} d\mathbf{x}$. Integrability of $\mathbf{q} \mathbf{b}$ is Jensen's inequality for the conditional expectation.

The norm is convex, so the conditional Jensen inequality gives $\|\mathbf{b}(\mathbf{I}_t)\| = \|\mathbb{E}[\dot{\mathbf{I}}_t \mid \mathbf{I}_t]\| \leq \mathbb{E}[\|\dot{\mathbf{I}}_t\| \mid \mathbf{I}_t]$ almost surely. Since $\mathbf{q}$ is the density of $\mathbf{I}_t$ (proof of [Proposition 3.3](#)), $\int \mathbf{q} \|\mathbf{b}\| d\mathbf{x} = \mathbb{E}\|\mathbf{b}(\mathbf{I}_t)\| \leq \mathbb{E}\|\dot{\mathbf{I}}_t\| < \infty$ by [Lemma B.3\(i\)](#).

Fix $\phi \in C_c^\infty$ and $(\mathbf{x}_0, \mathbf{x}_1)$. On $[0, 1]$, $t \mapsto \mathbf{I}_t$ is C^1 with derivative $\dot{\mathbf{I}}_t = \dot{\mathbf{A}}(t)(\mathbf{x}_0 - \mathbf{x}_1)$, so $t \mapsto \phi(\mathbf{I}_t)$ is C^1 on $[0, 1]$ with derivative $\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t$, and for $t' < 1$

$$\phi(\mathbf{I}_{t'}) - \phi(\mathbf{x}_0) = \int_0^{t'} \nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t dt,$$

using $\mathbf{I}_0 = \mathbf{x}_0$. The integrand is bounded by $\|\nabla \phi\|_\infty \sum_{\ell, k} |\dot{\alpha}_k(t)| (\|\mathbf{x}_0^{\ell, k}\| + 1)$, which is integrable on $(0, 1)$ (proof of [Lemma B.3\(i\)](#)). Letting $t' \uparrow 1$, $\mathbf{I}_{t'} \rightarrow \mathbf{I}_1 = \mathbf{x}_1$ since $\alpha_k(t') \rightarrow 0$, and dominated convergence on the right gives $\int_0^1 \nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t dt = \phi(\mathbf{x}_1) - \phi(\mathbf{x}_0)$. Taking expectations over $(\mathbf{x}_0, \mathbf{x}_1)$, and using that t is uniform on $[0, 1]$ and independent of $(\mathbf{x}_0, \mathbf{x}_1)$ together with Fubini's theorem (applicable because $|\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t| \leq \|\nabla \phi\|_\infty \|\dot{\mathbf{I}}_t\|$ is integrable by [Lemma B.3\(i\)](#)),

$$\mathbb{E}[\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t] = \mathbb{E} \int_0^1 \nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t dt = \mathbb{E}\phi(\mathbf{x}_1) - \mathbb{E}\phi(\mathbf{x}_0) = \int \phi d\mu_1 - \int \phi d\mu_0.$$

Since $\nabla \phi(\mathbf{I}_t)$ is bounded and $\sigma(\mathbf{I}_t)$ -measurable and $\dot{\mathbf{I}}_t$ is integrable, the tower property and [Eq. \(1\)](#) give $\mathbb{E}[\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t] = \mathbb{E}[\nabla \phi(\mathbf{I}_t) \cdot \mathbb{E}[\dot{\mathbf{I}}_t \mid \mathbf{I}_t]] = \mathbb{E}[\nabla \phi(\mathbf{I}_t) \cdot \mathbf{b}(\mathbf{I}_t)] = \int \nabla \phi \cdot \mathbf{b} \mathbf{q} d\mathbf{x}$. $\square$

We use the following decomposition theorem for normal one-dimensional currents.

Theorem B.7 (Smirnov [43]; Paolini–Stepanov [34, Prop. 3.8, Thms. 4.2 and 5.1]). *Let $\mathbf{T}$ be a normal one-dimensional current in $\mathbb{R}^D$ (finite mass $\mathbb{M}(\mathbf{T})$ and finite boundary mass). Then:*

- (i) *there is a cycle $\mathbf{C}$ of $\mathbf{T}$, i.e. $\partial \mathbf{C} = 0$ and $\mathbb{M}(\mathbf{T}) = \mathbb{M}(\mathbf{C}) + \mathbb{M}(\mathbf{T} - \mathbf{C})$, such that $\mathbf{A} := \mathbf{T} - \mathbf{C}$ is acyclic; in particular $\partial \mathbf{A} = \partial \mathbf{T}$ and $\|\mathbf{T}\| = \|\mathbf{A}\| + \|\mathbf{C}\|$;*
- (ii) *every acyclic normal one-dimensional current $\mathbf{A}$ is decomposable in curves: there is a finite Borel measure η on the space Θ of (equivalence classes of) Lipschitz curves, concentrated on arcs (injective curves), with*

$$\mathbf{A} = \int_{\Theta} [\theta] d\eta(\theta), \quad \mathbb{M}(\mathbf{A}) = \int_{\Theta} \ell_{\theta} d\eta(\theta), \quad (\mathbf{e})_{\#} \eta = (\partial \mathbf{A})^+, \quad (\mathbf{s})_{\#} \eta = (\partial \mathbf{A})^-,$$

where $\mathbf{s}, \mathbf{e}$ are the start and end points of a curve and $(\partial \mathbf{A})^{\pm}$ the positive and negative parts of the measure $\partial \mathbf{A}$; moreover $\|\mathbf{A}\| \leq \int_{\Theta} \|\theta\| d\eta(\theta)$.

[Theorem B.7](#) is stated in [34] for Ambrosio–Kirchheim currents in complete metric spaces; $\mathbb{R}^D$ is complete and separable, so no set-theoretic assumption is needed, and in $\mathbb{R}^D$ normal metric currents coincide with classical normal currents. We parametrise each arc by arc length, $\theta : [0, \ell_{\theta}] \rightarrow \mathbb{R}^D$ with $\|\theta'\| = 1$ a.e.; then $[\theta](\omega) = \int_0^{\ell_{\theta}} \omega(\theta(l)) \cdot \theta'(l) dl$, $\mathbb{M}([\theta]) = \ell_{\theta}$ and $\|[\theta]\| = \theta_{\#}(\mathcal{L}^1 \llcorner [0, \ell_{\theta}])$.

For $\mathbf{x} \in \Omega$ we write $[0, \tau(\mathbf{x}))$, $\tau(\mathbf{x}) \in (0, \infty]$, for the maximal interval of existence in Ω of the solution $\mathbf{X}_s(\mathbf{x})$ of $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$, $\mathbf{X}_0 = \mathbf{x}$ (unique, as $\mathbf{b}$ is locally Lipschitz on Ω by [Lemma 3.6](#)), as in the proof of [Theorem 3.7](#). We set $\tau := 0$ on $\mathcal{M}_1^+$. The set $\{(\mathbf{x}, s) : \mathbf{x} \in \Omega, 0 \leq s < \tau(\mathbf{x})\}$ is relatively open in $\Omega \times [0, \infty)$ and $(\mathbf{x}, s) \mapsto \mathbf{X}_s(\mathbf{x})$ is continuous on it (continuous dependence on initial data); in particular τ is lower semicontinuous on Ω , hence Borel.

Lemma B.8. *Let $\mathbf{x} \in \Omega$. Then $\|\mathbf{X}_s^{\ell,k}(\mathbf{x})\| \leq \max\{1, \|\mathbf{x}^{\ell,k}\|\}$ for all $s < \tau(\mathbf{x})$ and all blocks. If $\tau(\mathbf{x}) < \infty$, then $\lim_{s \uparrow \tau(\mathbf{x})} \mathbf{X}_s(\mathbf{x})$ exists and lies in $\mathcal{M}_1^+$.*

Proof. Idea. Two things are needed: trajectories stay bounded, and a trajectory that ceases to exist in finite time must end at a single point of $\mathcal{M}_1^+$. Boundedness follows from Eq. (7): each block is pulled towards a point of its simplex, which lies in the unit ball, so $\|\mathbf{X}_s^{\ell,k}\|^2$ decreases while it exceeds one. For the second claim, a maximal solution with finite existence time leaves every compact subset of Ω , so within the bounded region it must approach $\mathcal{M}_1^+$; as the points of $\mathcal{M}_1^+$ are at mutual distance at least $\sqrt{2}$, connectedness of the trajectory forces it to approach one of them.

Fix a block and let $u(s) := \|\mathbf{X}_s^{\ell,k}\|^2$. By Eq. (7), $\|\bar{\mathbf{p}}^{\ell,k}\| \leq \|\bar{\mathbf{p}}^{\ell,k}\|_1 = 1$ ($\bar{\mathbf{p}}^{\ell,k} \in \Delta^{v_k-1}$) and the Cauchy-Schwarz inequality,

$$u' = 2\langle \mathbf{X}_s^{\ell,k}, \mathbf{b}^{\ell,k}(\mathbf{X}_s) \rangle = 2a_k M(\mathbf{X}_s) (\langle \mathbf{X}_s^{\ell,k}, \bar{\mathbf{p}}^{\ell,k}(\mathbf{X}_s) \rangle - u) \leq 2a_k M(\mathbf{X}_s) (\sqrt{u} - u),$$

which is negative whenever $u > 1$ (as $M \geq 1$). Let $R := \max\{1, u(0)\}$ and suppose $u(s) > R$ for some s . With $s' := \sup\{r \leq s : u(r) \leq R\}$ we have $u(s') = R$ by continuity and $u > R \geq 1$ on $(s', s]$, so u is strictly decreasing on $[s', s]$ and $u(s) < u(s') = R$, a contradiction. This proves the first claim.

Now let $\tau(\mathbf{x}) < \infty$ and let $B := \{\mathbf{y} : \|\mathbf{y}^{\ell,k}\| \leq \max\{1, \|\mathbf{x}^{\ell,k}\|\} \forall \ell, k\}$, a compact set containing the trajectory. Distinct points of $\mathcal{M}_1^+$ differ in at least one block, where they are distinct vertices of a simplex, so their distance is at least $\sqrt{2}$. For $\delta \in (0, \sqrt{2}/2)$ the set $K_\delta := \{\mathbf{y} \in B : \text{dist}(\mathbf{y}, \mathcal{M}_1^+) \geq \delta\}$ is a compact subset of Ω , so (as in the proof of Theorem 3.7) there is $s_\delta < \tau(\mathbf{x})$ with $\mathbf{X}_s \notin K_\delta$, i.e. $\text{dist}(\mathbf{X}_s, \mathcal{M}_1^+) < \delta$, for $s \in (s_\delta, \tau(\mathbf{x}))$. The open balls of radius δ around the points of $\mathcal{M}_1^+$ are pairwise disjoint, and the connected set $\mathbf{X}((s_\delta, \tau(\mathbf{x})))$ lies in their union, hence in a single ball $B_\delta(\mathbf{e}_\delta)$. For $\delta' < \delta$ and $s > \max\{s_\delta, s_{\delta'}\}$, $\mathbf{X}_s \in B_\delta(\mathbf{e}_\delta) \cap B_{\delta'}(\mathbf{e}_{\delta'}) \subset B_\delta(\mathbf{e}_\delta) \cap B_{\delta'}(\mathbf{e}_{\delta'})$, so $\mathbf{e}_{\delta'} = \mathbf{e}_\delta =: \mathbf{e}$. Hence $\|\mathbf{X}_s - \mathbf{e}\| < \delta$ for $s > s_\delta$, for every small δ , i.e. $\mathbf{X}_s \rightarrow \mathbf{e} \in \mathcal{M}_1^+$. $\square$

Proof of Theorem 3.8. Idea. Theorem 3.7 says nothing about trajectories that start far from the data, and the flow cannot be integrated, so we argue with measures instead of single trajectories. By Lemma B.6 the current $q\mathbf{b}$ has boundary $\mu_1 - \mu_0$, and Theorem B.7 writes its acyclic part as a superposition of curves whose starting points have law μ_0 and whose end points have law μ_1 ; this is the right tool because it produces paths from noise to data out of the conservation law alone. It remains to show that these curves are reparametrised trajectories of $\mathbf{b}$ traversed in finite time: Steps 1–2 set up the decomposition, Step 3 shows that the curves are tangent to $\mathbf{b}$ because the current is, Step 4 bounds the expected travel time by $\int q = 1$, Step 5 identifies the curves with trajectories and uses Theorem 3.7 to exclude continuation through a vertex, and Steps 6–7 read off the four claims.

Step 1: the current. Let $\mathbf{T}(\omega) := \int \omega \cdot \mathbf{j} d\mathbf{x}$ with $\mathbf{j} = q\mathbf{b}$. By Lemma B.6, $\mathbb{M}(\mathbf{T}) = \sup_{\|\omega\|_\infty \leq 1} \mathbf{T}(\omega) = \int q\|\mathbf{b}\| < \infty$ and, for $\phi \in C_c^\infty$, $\partial \mathbf{T}(\phi) = \mathbf{T}(\nabla \phi) = \mu_1(\phi) - \mu_0(\phi)$, so $\partial \mathbf{T} = \mu_1 - \mu_0$. As μ_0 and μ_1 are mutually singular probability measures (Step 2), $\partial \mathbf{T}$ has mass $\mu_0(\mathbb{R}^D) + \mu_1(\mathbb{R}^D) = 2$, and $\mathbf{T}$ is normal. Apply Theorem B.7. A current of finite mass is given by a finite $\mathbb{R}^D$ -valued Borel measure (Riesz representation), and we use this measure to evaluate $\mathbf{A}$ and $\mathbf{C} = \mathbf{T} - \mathbf{A}$ on bounded Borel forms. The right-hand side of $\mathbf{A} = \int [\theta] d\eta$ is likewise integration against the vector measure $B \mapsto \int_\Theta \int_0^{\ell_\theta} \mathbb{K}_B(\theta(l)) \theta'(l) dl d\eta(\theta)$, whose total variation is at most $\int \ell_\theta d\eta = \mathbb{M}(\mathbf{A}) < \infty$. Two finite vector measures that agree on C_c^∞ forms agree (by uniform density of C_c^∞ in C_0 and uniqueness in the Riesz representation), so the identity $\mathbf{A}(\omega) = \int_\Theta [\theta](\omega) d\eta(\theta)$ extends from smooth to bounded Borel forms.

Step 2: start and end points. $\partial \mathbf{A} = \partial \mathbf{T} = \mu_1 - \mu_0$. The measure μ_0 has a density, so it vanishes on the finite set $\mathcal{M}_1^+$, which carries μ_1 ; hence $\mu_0 \perp \mu_1$, and by uniqueness of the Jordan decomposition $(\partial \mathbf{A})^+ = \mu_1$ and $(\partial \mathbf{A})^- = \mu_0$. Theorem B.7(ii) gives

$$s_\# \eta = \mu_0, \quad e_\# \eta = \mu_1; \quad (49)$$

in particular $\eta(\Theta) = \mu_0(\mathbb{R}^D) = 1$. We write η' for η below, to keep the notation of the following steps.

Step 3: the curves follow the field. Let $\mathcal{N} := \{\mathbf{b} = \mathbf{0}\} \cup \mathcal{M}_1^+$ and $\psi := \mathbf{b}/\|\mathbf{b}\|$ off $\mathcal{N}$, $\psi := \mathbf{0}$ on $\mathcal{N}$; ψ is Borel because $\mathbf{b}$ is continuous on Ω . The measure $\|\mathbf{T}\| = q\|\mathbf{b}\| d\mathbf{x}$ gives $\mathcal{N}$ zero mass ($q\|\mathbf{b}\| = 0$ on

$\{\mathbf{b} = \mathbf{0}\}$ and $\mathcal{M}_1^+$ is Lebesgue-null), so $\int \boldsymbol{\psi} \cdot d\mathbf{T} = \int_{\mathbb{R}^D \setminus \mathcal{N}} \frac{\mathbf{b}}{\|\mathbf{b}\|} \cdot \mathbf{b} q d\mathbf{x} = \int q \|\mathbf{b}\| = \mathbb{M}(\mathbf{T})$. Since $\|\boldsymbol{\psi}\| \leq 1$, $\int \boldsymbol{\psi} \cdot d\mathbf{A} \leq \mathbb{M}(\mathbf{A})$ and $\int \boldsymbol{\psi} \cdot d\mathbf{C} \leq \mathbb{M}(\mathbf{C})$, and by [Theorem B.7\(i\)](#)

$$\mathbb{M}(\mathbf{T}) = \int \boldsymbol{\psi} \cdot d\mathbf{A} + \int \boldsymbol{\psi} \cdot d\mathbf{C} \leq \mathbb{M}(\mathbf{A}) + \mathbb{M}(\mathbf{C}) = \mathbb{M}(\mathbf{T}),$$

so both inequalities are equalities; in particular $\int \boldsymbol{\psi} \cdot d\mathbf{A} = \mathbb{M}(\mathbf{A})$. Also $\|\mathbf{A}\| \leq \|\mathbf{A}\| + \|\mathbf{C}\| = \|\mathbf{T}\| = \int q \|\mathbf{b}\| d\mathbf{x}$ by [Theorem B.7\(i\)](#). Next, by Step 1 and $\boldsymbol{\psi} \cdot \boldsymbol{\theta}' \leq \|\boldsymbol{\psi}\| \|\boldsymbol{\theta}'\| \leq 1$,

$$\mathbb{M}(\mathbf{A}) = \int \boldsymbol{\psi} \cdot d\mathbf{A} = \int_{\Theta} \int_0^{\ell_{\theta}} \boldsymbol{\psi}(\theta(l)) \cdot \boldsymbol{\theta}'(l) dl d\eta(\theta) \leq \int_{\Theta} \ell_{\theta} d\eta = \mathbb{M}(\mathbf{A}).$$

Hence $\int_{\Theta} \int_0^{\ell_{\theta}} (1 - \boldsymbol{\psi}(\theta(l)) \cdot \boldsymbol{\theta}'(l)) dl d\eta = 0$ with a nonnegative integrand, so for η -a.e. θ : $\boldsymbol{\psi}(\theta(l)) \cdot \boldsymbol{\theta}'(l) = 1$ for a.e. l . As $\|\boldsymbol{\psi}\| \leq 1$ and $\|\boldsymbol{\theta}'\| = 1$, equality in the Cauchy–Schwarz inequality forces $\|\boldsymbol{\psi}(\theta(l))\| = 1$, i.e. $\theta(l) \notin \mathcal{N}$, and $\boldsymbol{\theta}'(l) = \boldsymbol{\psi}(\theta(l)) = \mathbf{b}(\theta(l))/\|\mathbf{b}(\theta(l))\|$, for a.e. l . Finally, $\|\mathbf{A}\| \leq \lambda := \int \|\boldsymbol{\theta}\| d\eta = \int \theta_{\#}(\mathcal{L}^1_{[0, \ell_{\theta}]}) d\eta$ by [Theorem B.7\(ii\)](#), and the total masses agree: $\|\mathbf{A}\|(\mathbb{R}^D) = \mathbb{M}(\mathbf{A}) = \int \ell_{\theta} d\eta = \lambda(\mathbb{R}^D)$. So $\lambda - \|\mathbf{A}\|$ is a nonnegative measure of total mass 0, i.e. $\|\mathbf{A}\| = \lambda$. For Borel $g \geq 0$, $\int g d\lambda = \int_{\Theta} \int_0^{\ell_{\theta}} g(\theta(l)) dl d\eta(\theta)$ (by definition for indicators, then for simple functions by linearity and for $g \geq 0$ by monotone convergence), and with $\|\mathbf{A}\| \leq \int q \|\mathbf{b}\| d\mathbf{x}$,

$$\int_{\Theta} \int_0^{\ell_{\theta}} g(\theta(l)) dl d\eta(\theta) = \int g d\|\mathbf{A}\| \leq \int g q \|\mathbf{b}\| d\mathbf{x}. \quad (50)$$

Step 4: finite time along the curves. Define the travel time $\mathsf{T}(\theta) := \int_0^{\ell_{\theta}} dl / \|\mathbf{b}(\theta(l))\|$ (with $1/\|\mathbf{b}\| := 0$ on $\mathcal{N}$, irrelevant by Step 3). By [Eq. \(50\)](#) with the Borel function $g = \mathbb{1}_{\mathbb{R}^D \setminus \mathcal{N}} / \|\mathbf{b}\| \geq 0$, and since q is a probability density,

$$\int_{\Theta} \mathsf{T} d\eta = \int_{\Theta} \int_0^{\ell_{\theta}} g(\theta(l)) dl d\eta(\theta) \leq \int_{\mathbb{R}^D \setminus \mathcal{N}} \frac{q \|\mathbf{b}\|}{\|\mathbf{b}\|} d\mathbf{x} \leq \int_{\mathbb{R}^D} q d\mathbf{x} = 1.$$

Hence $\mathsf{T}(\theta) < \infty$ for η -a.e. θ . Let $\mathcal{G} \subset \Theta$ be the set of arcs (injective curves) with $\mathsf{T}(\theta) < \infty$, the properties of Step 3, $\theta(0) \notin \mathcal{M}_1^+$ and $\theta(\ell_{\theta}) \in \mathcal{M}_1^+$. Each of these properties holds for η -a.e. θ : injectivity by [Theorem B.7\(ii\)](#), the next two by Steps 3–4, and the last two by [Eq. \(49\)](#), since $\eta(s \in \mathcal{M}_1^+) = \mu_0(\mathcal{M}_1^+) = 0$ and $\eta(e \notin \mathcal{M}_1^+) = 1 - \mu_1(\mathcal{M}_1^+) = 0$. Hence $\eta'(\Theta \setminus \mathcal{G}) = 0$.

Step 5: curves are trajectories. Fix $\theta \in \mathcal{G}$, $\mathbf{x} := \theta(0)$, $g(l) := 1/\|\mathbf{b}(\theta(l))\|$ (with the convention of Step 4), and $\mathbf{r}(l) := \int_0^l g$. Then $g \in L^1(0, \ell_{\theta})$ (as $\mathsf{T}(\theta) < \infty$), $\mathbf{r}$ is absolutely continuous with $\mathbf{r}' = g$ a.e., and $\mathbf{r}$ is strictly increasing because $g > 0$ a.e. ($\theta(l) \notin \mathcal{N}$ for a.e. l), so $\mathbf{r}(l') - \mathbf{r}(l) = \int_l^{l'} g > 0$ for $l < l'$; moreover $\mathbf{r}(\ell_{\theta}) = \mathsf{T}(\theta)$. Let $\lambda := \mathbf{r}^{-1} : [0, \mathsf{T}(\theta)] \rightarrow [0, \ell_{\theta}]$ (continuous, strictly increasing) and $\gamma(r) := \theta(\lambda(r))$, so that $\gamma(\mathbf{r}(l)) = \theta(l)$. The set $\{r \in [0, \mathsf{T}(\theta)] : \gamma(r) \in \mathcal{M}_1^+\}$ is closed (γ is continuous and $\mathcal{M}_1^+$ is finite) and contains $\mathsf{T}(\theta)$, as $\gamma(\mathsf{T}(\theta)) = \theta(\ell_{\theta}) \in \mathcal{M}_1^+$; let $r_0 \leq \mathsf{T}(\theta)$ be its minimum. Then $r_0 > 0$ as $\gamma(0) = \mathbf{x} \notin \mathcal{M}_1^+$, and $\gamma(r) \in \Omega$ for $r < r_0$. Fix $r < r_0$. The set $\gamma([0, r])$ is a compact subset of Ω , on which the continuous field $\mathbf{b}$ is bounded, so $\mathbf{h}(r') := \mathbf{b}(\gamma(r'))$ is bounded and continuous on $[0, r]$. Since θ is Lipschitz, hence absolutely continuous, and by Step 3 $\boldsymbol{\theta}'(l) = \mathbf{b}(\theta(l)) g(l) = \mathbf{h}(\mathbf{r}(l)) \mathbf{r}'(l)$ for a.e. l ,

$$\gamma(r) - \mathbf{x} = \int_0^{\lambda(r)} \boldsymbol{\theta}'(l) dl = \int_0^{\lambda(r)} \mathbf{b}(\theta(l)) g(l) dl = \int_0^{\lambda(r)} \mathbf{h}(\mathbf{r}(l)) \mathbf{r}'(l) dl = \int_0^r \mathbf{h}(r') dr',$$

where the last step is the change-of-variables formula for the absolutely continuous increasing map $\mathbf{r}$ [\[41\]](#), applied to the positive and negative parts of each coordinate of $\mathbf{h}$ (note $\|\mathbf{h}(\mathbf{r}(l)) \mathbf{r}'(l)\| = 1$ a.e.), and $\mathbf{r}(0) = 0$, $\mathbf{r}(\lambda(r)) = r$. The integrand $\mathbf{h} = \mathbf{b} \circ \gamma$ is continuous on $[0, r_0)$, so γ is C^1 on $[0, r_0)$ with $\gamma' = \mathbf{b}(\gamma)$ and $\gamma(0) = \mathbf{x}$. Since $\mathbf{b}$ is locally Lipschitz on Ω , uniqueness gives $\gamma = \mathbf{X}_{\cdot}(\mathbf{x})$ on $[0, r_0)$ and $\tau(\mathbf{x}) \geq r_0$. If $\tau(\mathbf{x}) > r_0$, then $\mathbf{X}_{r_0}(\mathbf{x}) \in \Omega$ and, by continuity, $\mathbf{X}_{r_0}(\mathbf{x}) = \lim_{r \uparrow r_0} \gamma(r) = \gamma(r_0) \in \mathcal{M}_1^+$, a contradiction. Hence the maximal solution in Ω is defined exactly on $[0, r_0)$: $\tau(\mathbf{x}) = r_0 < \infty$ and $\mathbf{T}(\mathbf{x}) = \lim_{r \uparrow r_0} \gamma(r) = \gamma(r_0)$.

No continuation through a vertex. Suppose $r_0 < \mathsf{T}(\theta)$ and let $\mathbf{e}_{j^*} := \gamma(r_0)$, with ε from [Theorem 3.7](#). Since $V_{j^*}(\gamma(r_0)) = 0$ and $V_{j^*} \circ \gamma$ is continuous, there is $\rho \in (0, \mathsf{T}(\theta) - r_0]$ with $\gamma([r_0, r_0 + \rho]) \subset \{V_{j^*} < \varepsilon\}$. As θ is injective and λ strictly increasing, γ is injective, so $\gamma(r) \neq \mathbf{e}_{j^*}$ for $r \neq r_0$; by [Eq. \(47\)](#),

$\gamma((r_0, r_0 + \rho]) \subset W = \{0 < V_{j^*} < \varepsilon\} \subset \Omega$. For $r_1 \in (r_0, r_0 + \rho)$, the computation above on $[\lambda(r_1), \lambda(r)]$ gives $\gamma(r) - \gamma(r_1) = \int_{r_1}^r \mathbf{b}(\gamma(r')) dr'$ for $r \in [r_1, r_0 + \rho]$, so γ is an integral curve of $\mathbf{b}$ on $[r_1, r_0 + \rho]$ starting in W ; by uniqueness and Theorem 3.7(i), $r \mapsto V_{j^*}(\gamma(r))$ is decreasing on $[r_1, r_0 + \rho]$. As r_1 is arbitrary, $V_{j^*}(\gamma(r')) \leq V_{j^*}(\gamma(r))$ for $r_0 < r < r' \leq r_0 + \rho$. Letting $r \downarrow r_0$, $V_{j^*}(\gamma(r)) \rightarrow V_{j^*}(\mathbf{e}_{j^*}) = 0$, so $V_{j^*}(\gamma(r')) \leq 0$, i.e. $\gamma(r') = \mathbf{e}_{j^*} = \gamma(r_0)$, contradicting injectivity. Hence $r_0 = T(\theta)$ and, as $\gamma(T(\theta)) = \theta(\ell_\theta)$,

$$\tau(\theta(0)) = T(\theta), \quad T(\theta(0)) = \theta(\ell_\theta) \quad (\theta \in \mathcal{G}). \quad (51)$$

Step 6: openness and local constancy. Let $\mathbf{x} \in \mathcal{A}$. By Lemma B.8, $T(\mathbf{x})$ exists and equals some $\mathbf{e}_{j^*} \in \mathcal{M}_1^+$. Since $\mathbf{X}_s(\mathbf{x}) \rightarrow \mathbf{e}_{j^*}$ and $\mathbf{X}_s(\mathbf{x}) \in \Omega$, there is $s_1 < \tau(\mathbf{x})$ with $\mathbf{X}_{s_1}(\mathbf{x}) \in W = \{0 < V_{j^*} < \varepsilon\}$ (ε and W from the proof of Theorem 3.7). The trajectory segment $\mathbf{X}_{[0, s_1]}(\mathbf{x})$ is a compact subset of Ω , so by continuous dependence on initial data the map $\mathbf{y} \mapsto \mathbf{X}_{s_1}(\mathbf{y})$ is defined and continuous on a neighbourhood $U \subset \Omega$ of $\mathbf{x}$, and since W is open, shrinking U we get $\mathbf{X}_{s_1}(\mathbf{y}) \in W$ for $\mathbf{y} \in U$. By Theorem 3.7(i) applied at $\mathbf{X}_{s_1}(\mathbf{y})$ and the semigroup property $\mathbf{X}_{s_1+s}(\mathbf{y}) = \mathbf{X}_s(\mathbf{X}_{s_1}(\mathbf{y}))$, every $\mathbf{y} \in U$ has $\tau(\mathbf{y}) = s_1 + \tau(\mathbf{X}_{s_1}(\mathbf{y})) < \infty$ and $T(\mathbf{y}) = \mathbf{e}_{j^*}$. So $U \subset \mathcal{A}$, $\mathcal{A}$ is open, and T is constant on a neighbourhood of each point of $\mathcal{A}$; hence $T \in C^\infty(\mathcal{A})$ with $\mathbf{J}_T = \mathbf{0}$, and $\mathbf{J}_T \mathbf{b} = \mathbf{0}$ on $\mathcal{A}$.

Step 7: conclusions. $\mathcal{A}$ is open, hence Borel, and $s^{-1}(\mathcal{A}) \supset \mathcal{G}$ by Eq. (51) (for $\theta \in \mathcal{G}$, $\theta(0) \in \Omega$ and $\tau(\theta(0)) = T(\theta) < \infty$), so by Eq. (49) and Step 4, $\mu_0(\mathcal{A}) = \eta'(s^{-1}(\mathcal{A})) \geq \eta'(\mathcal{G}) = 1$. As μ_0 has a strictly positive density ϱ , $0 = \mu_0(\mathbb{R}^D \setminus \mathcal{A}) = \int_{\mathbb{R}^D \setminus \mathcal{A}} \varrho d\mathbf{x}$ forces $\mathbb{R}^D \setminus \mathcal{A}$ to be Lebesgue-null; with Step 6 and Lemma B.8 this proves the first part of (i) and (ii). By Eq. (51), $T \circ s = e$ and $\tau \circ s = T$ on $\mathcal{G}$, hence η' -a.e. (T is Borel on $\mathcal{A}$, being locally constant; we extend it by a constant on the μ_0 -null set $\mathbb{R}^D \setminus \mathcal{A}$). Therefore

$$T_{\#} \mu_0 = T_{\#} s_{\#} \eta' = (T \circ s)_{\#} \eta' = e_{\#} \eta' = \mu_1,$$

which is (iii), and $\mathbb{E}_{\mu_0} \tau = \int \tau \circ s d\eta' = \int T d\eta' \leq 1$ by Step 4. For (iv), fix Borel $\phi \geq 0$. The map $\mathbf{x} \mapsto \int_0^{\tau(\mathbf{x})} \phi(\mathbf{X}_r(\mathbf{x})) dr$ is Borel on Ω by Tonelli's theorem, as $(\mathbf{x}, r) \mapsto \phi(\mathbf{X}_r(\mathbf{x})) \mathbb{1}_{\{r < \tau(\mathbf{x})\}}$ is Borel (the flow is continuous on its relatively open domain). For $\theta \in \mathcal{G}$, $\mathbf{X}_r(\theta(0)) = \gamma(r)$ for $r < \tau(\theta(0)) = T(\theta)$ (Step 5), and the substitution $r = r(l)$ of Step 5 (change of variables for the nonnegative Borel function $\phi \circ \gamma$ [41]) gives $\int_0^{\tau(\theta(0))} \phi(\mathbf{X}_r(\theta(0))) dr = \int_0^{\ell_\theta} \phi(\theta(l)) g(l) dl$. Integrating over η' , using $s_{\#} \eta' = \mu_0$ and $\eta'(\Theta \setminus \mathcal{G}) = 0$, and applying Eq. (50) with the Borel function $\phi \mathbb{1}_{\mathbb{R}^D \setminus \mathcal{N}} / \|\mathbf{b}\|$,

$$\mathbb{E}_{\mu_0} \int_0^\tau \phi(\mathbf{X}_r) dr = \int_{\mathcal{G}} \int_0^{\ell_\theta} \frac{\phi(\theta(l))}{\|\mathbf{b}(\theta(l))\|} dl d\eta'(\theta) \leq \int_{\mathbb{R}^D \setminus \mathcal{N}} \frac{\phi q \|\mathbf{b}\|}{\|\mathbf{b}\|} d\mathbf{x} \leq \int \phi q d\mathbf{x}. \quad \square$$

C Proofs for Section 4

C.1 Proof of Proposition 4.1

Proof. Idea. The posterior is prior times likelihood divided by a normaliser, and prior and likelihood are both products over blocks, so the numerator factorises at once. What must be shown is that the normaliser, a sum over the product set $\mathcal{J}$, is also a product, and that each factor is the posterior of a single block: the distributive law of Step 4 gives the first, and integrating out the other blocks in Step 6 gives the second.

Write $\mathcal{B} := [L] \times [K]$ for the set of blocks and $b = (\ell, k) \in \mathcal{B}$ for a block, with size $n_b := v_k$, schedule value $\alpha_b := \alpha_k(t)$ and prior scale $\sigma_b := \sigma_k$; let $N := |\mathcal{B}| = LK$ and fix an enumeration $b_1, \dots, b_N$ of $\mathcal{B}$. The ambient space is the orthogonal direct sum

$$\mathbb{R}^D = \bigoplus_{b \in \mathcal{B}} \mathbb{R}^{n_b}, \quad \mathbf{x} = \bigoplus_{b \in \mathcal{B}} \mathbf{x}^b, \quad \langle \mathbf{x}, \mathbf{y} \rangle = \sum_{b \in \mathcal{B}} \langle \mathbf{x}^b, \mathbf{y}^b \rangle, \quad \|\mathbf{x}\|^2 = \sum_{b \in \mathcal{B}} \|\mathbf{x}^b\|^2, \quad (52)$$

the configurations form the product set $\mathcal{J} = \prod_{b \in \mathcal{B}} [n_b]$, and the vertex of $\mathbf{j} = (j_b)_b$ is $\mathbf{e}_{\mathbf{j}} = \bigoplus_b \mathbf{e}_{j_b}$. The hypothesis $\mu_1 = \bigotimes_b \mu_1^b$ with $\mu_1^b = \sum_{i=1}^{n_b} \pi_i^b \delta_{\mathbf{e}_i}$ means

$$\pi_{\mathbf{j}} = \mu_1(\{\mathbf{e}_{\mathbf{j}}\}) = \left(\bigotimes_{b \in \mathcal{B}} \mu_1^b \right) \left(\prod_{b \in \mathcal{B}} \{\mathbf{e}_{j_b}\} \right) = \prod_{b \in \mathcal{B}} \mu_1^b(\{\mathbf{e}_{j_b}\}) = \prod_{b \in \mathcal{B}} \pi_{j_b}^b \quad (\mathbf{j} \in \mathcal{J}). \quad (53)$$

Step 1: Bayes' rule at fixed t . Since t is independent of $\mathbf{x}_1$, conditioning on t does not change the law of $\mathbf{x}_1$, and the conditional density of $\mathbf{I}_t$ given $(t, \mathbf{x}_1 = \mathbf{e}_j)$ is $p(\cdot | t, \mathbf{j})$ (Lemma B.1). For every $\mathbf{x} \in \mathbb{R}^D$,

$$\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j | \mathbf{I}_t = \mathbf{x}, t) = \frac{\pi_j p(\mathbf{x} | t, \mathbf{j})}{Z(\mathbf{x})}, \quad Z(\mathbf{x}) := \sum_{j' \in \mathcal{J}} \pi_{j'} p(\mathbf{x} | t, \mathbf{j}'). \quad (54)$$

Step 2: the likelihood is a tensor product of block likelihoods. The conditional covariance is the direct sum $\mathbf{A}\Sigma\mathbf{A} = \bigoplus_b \alpha_b^2 \sigma_b^2 \mathbf{I}_{n_b}$, and the conditional mean is $(\mathbf{I} - \mathbf{A})\mathbf{e}_j = \bigoplus_b (1 - \alpha_b) \mathbf{e}_{j_b}$. For a direct sum of matrices the determinant and the inverse act blockwise, $\det \bigoplus_b M_b = \prod_b \det M_b$ and $(\bigoplus_b M_b)^{-1} = \bigoplus_b M_b^{-1}$, so with $\mathbf{r} := \mathbf{x} - (\mathbf{I} - \mathbf{A})\mathbf{e}_j = \bigoplus_b \mathbf{r}^b$, $\mathbf{r}^b = \mathbf{x}^b - (1 - \alpha_b) \mathbf{e}_{j_b}$,

$$\begin{aligned} p(\mathbf{x} | t, \mathbf{j}) &= (2\pi)^{-D/2} \left(\det \bigoplus_b \alpha_b^2 \sigma_b^2 \mathbf{I}_{n_b} \right)^{-1/2} \exp \left(-\frac{1}{2} \left\langle \mathbf{r}, \left(\bigoplus_b \alpha_b^2 \sigma_b^2 \mathbf{I}_{n_b} \right)^{-1} \mathbf{r} \right\rangle \right) \\ &= (2\pi)^{-\sum_b n_b/2} \left(\prod_b (\alpha_b^2 \sigma_b^2)^{n_b} \right)^{-1/2} \exp \left(-\frac{1}{2} \sum_b \langle \mathbf{r}^b, (\alpha_b^2 \sigma_b^2)^{-1} \mathbf{r}^b \rangle \right) \\ &= \prod_b (2\pi)^{-n_b/2} (\alpha_b^2 \sigma_b^2)^{-n_b/2} \prod_b \exp \left(-\frac{\|\mathbf{r}^b\|^2}{2\alpha_b^2 \sigma_b^2} \right) = \prod_{b \in \mathcal{B}} p^b(\mathbf{x}^b | t, j_b), \end{aligned} \quad (55)$$

where $D = \sum_b n_b$, the inner product was split by Eq. (52), exp turns the sum into a product, and

$$p^b(\mathbf{y} | t, i) := (2\pi \alpha_b^2 \sigma_b^2)^{-n_b/2} \exp \left(-\frac{\|\mathbf{y} - (1 - \alpha_b) \mathbf{e}_i\|^2}{2\alpha_b^2 \sigma_b^2} \right), \quad \mathbf{y} \in \mathbb{R}^{n_b}, \quad i \in [n_b],$$

is the density of $\mathcal{N}((1 - \alpha_b) \mathbf{e}_i, \alpha_b^2 \sigma_b^2 \mathbf{I}_{n_b})$. In operator form, the conditional law of $\mathbf{I}_t$ is the tensor product $\bigotimes_b \mathcal{N}((1 - \alpha_b) \mathbf{e}_{j_b}, \alpha_b^2 \sigma_b^2 \mathbf{I}_{n_b})$ on $\bigoplus_b \mathbb{R}^{n_b}$.

Step 3: the numerator. By Eqs. (53) and (55), with $f^b(i) := \pi_i^b p^b(\mathbf{x}^b | t, i) \geq 0$,

$$\pi_j p(\mathbf{x} | t, \mathbf{j}) = \left(\prod_b \pi_{j_b}^b \right) \left(\prod_b p^b(\mathbf{x}^b | t, j_b) \right) = \prod_{b \in \mathcal{B}} f^b(j_b). \quad (56)$$

Step 4: the distributive law over the product set. We claim that for any finite index sets and any functions $f^b : [n_b] \rightarrow \mathbb{R}$,

$$\sum_{j \in \prod_{b \in \mathcal{B}} [n_b]} \prod_{b \in \mathcal{B}} f^b(j_b) = \prod_{b \in \mathcal{B}} \sum_{i=1}^{n_b} f^b(i). \quad (57)$$

We prove this by induction on the number m of blocks, writing $\mathcal{J}_m = \prod_{r \leq m} [n_{b_r}]$. For $m = 1$ both sides are $\sum_i f^{b_1}(i)$. Suppose Eq. (57) holds for $b_1, \dots, b_{m-1}$. Every $\mathbf{j} \in \mathcal{J}_m$ is a unique pair $(\mathbf{j}_{<m}, j_{b_m}) \in \mathcal{J}_{m-1} \times [n_{b_m}]$, so the sum over $\mathcal{J}_m$ is an iterated sum, and

$$\begin{aligned} \sum_{j \in \mathcal{J}_m} \prod_{r=1}^m f^{b_r}(j_{b_r}) &= \sum_{\mathbf{j}_{<m} \in \mathcal{J}_{m-1}} \sum_{i=1}^{n_{b_m}} \left(\prod_{r=1}^{m-1} f^{b_r}(j_{b_r}) \right) f^{b_m}(i) \\ &= \sum_{\mathbf{j}_{<m} \in \mathcal{J}_{m-1}} \left(\prod_{r=1}^{m-1} f^{b_r}(j_{b_r}) \right) \left(\sum_{i=1}^{n_{b_m}} f^{b_m}(i) \right) \\ &= \left(\sum_{\mathbf{j}_{<m} \in \mathcal{J}_{m-1}} \prod_{r=1}^{m-1} f^{b_r}(j_{b_r}) \right) \left(\sum_{i=1}^{n_{b_m}} f^{b_m}(i) \right) = \left(\prod_{r=1}^{m-1} \sum_{i=1}^{n_{b_r}} f^{b_r}(i) \right) \left(\sum_{i=1}^{n_{b_m}} f^{b_m}(i) \right) \\ &= \prod_{r=1}^m \sum_{i=1}^{n_{b_r}} f^{b_r}(i), \end{aligned}$$

where the second and third equalities pull factors that do not depend on the summation variable out of the sums and the fourth is the induction hypothesis. Applied to the f^b of Step 3, Eqs. (56) and (57) give

$$Z(\mathbf{x}) = \sum_{j' \in \mathcal{J}} \prod_b f^b(j'_b) = \prod_{b \in \mathcal{B}} Z^b(\mathbf{x}^b), \quad Z^b(\mathbf{x}^b) := \sum_{i=1}^{n_b} \pi_i^b p^b(\mathbf{x}^b | t, i) > 0. \quad (58)$$

Step 5: the joint posterior is a product. Dividing Eq. (56) by Eq. (58) factor by factor,

$$\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j \mid \mathbf{I}_t = \mathbf{x}, t) = \frac{\prod_b f^b(j_b)}{\prod_b Z^b(\mathbf{x}^b)} = \prod_{b \in \mathcal{B}} q^b(j_b \mid \mathbf{x}^b), \quad q^b(i \mid \mathbf{x}^b) := \frac{\pi_i^b p^b(\mathbf{x}^b \mid t, i)}{\sum_{i'=1}^{n_b} \pi_{i'}^b p^b(\mathbf{x}^b \mid t, i')}, \quad (59)$$

i.e. the posterior law of $\mathbf{x}_1$ is the tensor product $\bigotimes_b \sum_i q^b(i \mid \mathbf{x}^b) \delta_{\mathbf{e}_i}$: given $\mathbf{I}_t$ and t , the blocks of $\mathbf{x}_1$ are independent, and block b depends on $\mathbf{x}$ only through $\mathbf{x}^b$.

Step 6: each factor is the block posterior. It remains to identify q^b with $\mathbb{P}(\mathbf{x}_1^b = \mathbf{e}_i \mid \mathbf{I}_t^b = \mathbf{x}^b, t)$. First, the marginal of $\mathbf{x}_1^b$ is μ_1^b : summing Eq. (53) over all configurations with $j_b = i$ and applying Eq. (57) to the remaining blocks,

$$\mathbb{P}(\mathbf{x}_1^b = \mathbf{e}_i) = \sum_{j: j_b = i} \prod_{b'} \pi_{j_{b'}}^{b'} = \pi_i^b \prod_{b' \neq b} \sum_{i'=1}^{n_{b'}} \pi_{i'}^{b'} = \pi_i^b.$$

Second, the conditional density of $\mathbf{I}_t^b$ given $(t, \mathbf{x}_1^b = \mathbf{e}_i)$ is $p^b(\cdot \mid t, i)$: by Bayes' rule and Eq. (55), integrating out the other blocks (Fubini, all integrands nonnegative),

$$\begin{aligned} \mathbb{P}(\mathbf{x}_1^b = \mathbf{e}_i) p_{\mathbf{I}_t^b \mid \mathbf{x}_1^b}(\mathbf{y} \mid i) &= \sum_{j: j_b = i} \pi_j \int_{\bigoplus_{b' \neq b} \mathbb{R}^{n_{b'}}} p(\mathbf{y} \oplus \mathbf{z} \mid t, \mathbf{j}) d\mathbf{z} \\ &= \sum_{j: j_b = i} \left(\prod_{b'} \pi_{j_{b'}}^{b'} \right) p^b(\mathbf{y} \mid t, i) \prod_{b' \neq b} \int_{\mathbb{R}^{n_{b'}}} p^{b'}(\mathbf{z}^{b'} \mid t, j_{b'}) d\mathbf{z}^{b'} \\ &= p^b(\mathbf{y} \mid t, i) \pi_i^b \sum_{(j_{b'})_{b' \neq b}} \prod_{b' \neq b} \pi_{j_{b'}}^{b'} \cdot 1 = p^b(\mathbf{y} \mid t, i) \pi_i^b \prod_{b' \neq b} \sum_{i'=1}^{n_{b'}} \pi_{i'}^{b'} = \pi_i^b p^b(\mathbf{y} \mid t, i), \end{aligned}$$

using that each Gaussian density integrates to one and Eq. (57) once more. Bayes' rule for the single block then gives $\mathbb{P}(\mathbf{x}_1^b = \mathbf{e}_i \mid \mathbf{I}_t^b = \mathbf{x}^b, t) = \pi_i^b p^b(\mathbf{x}^b \mid t, i) / \sum_{i'} \pi_{i'}^b p^b(\mathbf{x}^b \mid t, i') = q^b(i \mid \mathbf{x}^b)$, which proves the first claim.

Step 7: the softmax for uniform blocks. Let $\pi_i^b = 1/n_b$ for all i . Using $\|\mathbf{e}_i\|^2 = 1$ and $\langle \mathbf{x}^b, \mathbf{e}_i \rangle = x_i^b$,

$$\|\mathbf{x}^b - (1 - \alpha_b) \mathbf{e}_i\|^2 = \|\mathbf{x}^b\|^2 - 2(1 - \alpha_b) \langle \mathbf{x}^b, \mathbf{e}_i \rangle + (1 - \alpha_b)^2 \|\mathbf{e}_i\|^2 = \|\mathbf{x}^b\|^2 - 2(1 - \alpha_b) x_i^b + (1 - \alpha_b)^2,$$

so that

$$\pi_i^b p^b(\mathbf{x}^b \mid t, i) = \underbrace{\frac{1}{n_b} (2\pi\alpha_b^2\sigma_b^2)^{-n_b/2} \exp\left(-\frac{\|\mathbf{x}^b\|^2 + (1 - \alpha_b)^2}{2\alpha_b^2\sigma_b^2}\right)}_{=: c^b(\mathbf{x}^b), \text{ independent of } i} \cdot \exp\left(\frac{2(1 - \alpha_b) x_i^b}{2\alpha_b^2\sigma_b^2}\right) = c^b(\mathbf{x}^b) e^{\beta_b x_i^b},$$

with $\beta_b = (1 - \alpha_b)/(\alpha_b^2\sigma_b^2) = \beta_k(t)$. Substituting into Eq. (59), the factor $c^b(\mathbf{x}^b) > 0$ cancels between numerator and denominator:

$$q^b(i \mid \mathbf{x}^b) = \frac{c^b(\mathbf{x}^b) e^{\beta_b x_i^b}}{\sum_{i'=1}^{n_b} c^b(\mathbf{x}^b) e^{\beta_b x_{i'}^b}} = \frac{e^{\beta_b x_i^b}}{\sum_{i'=1}^{n_b} e^{\beta_b x_{i'}^b}} = \text{softmax}(\beta_b \mathbf{x}^b)_i,$$

which is Eq. (10). Combining with Eq. (59), the full posterior in log form reads

$$\log \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j \mid \mathbf{I}_t = \mathbf{x}, t) = \sum_{(\ell, k) \in \mathcal{B}} \left(\beta_k(t) x_{j_{\ell, k}}^{\ell, k} - \log \sum_{i=1}^{v_k} e^{\beta_k(t) x_i^{\ell, k}} \right).$$

□

Remark C.1 (The posterior entropy is additive). For a product law $q(\mathbf{j}) = \prod_b q^b(j_b)$ on $\mathcal{J}$, $\log q(\mathbf{j}) = \sum_b \log q^b(j_b)$, so by linearity and Eq. (57) (with all but one factor summing to one),

$$H(q) = - \sum_{\mathbf{j} \in \mathcal{J}} \left(\prod_{b'} q^{b'}(j_{b'}) \right) \sum_b \log q^b(j_b) = - \sum_b \sum_{i=1}^{n_b} q^b(i) \log q^b(i) \prod_{b' \neq b} \sum_{i'=1}^{n_{b'}} q^{b'}(i') = \sum_{b \in \mathcal{B}} H(q^b).$$

Under the hypotheses of Proposition 4.1, $H(\mathbf{x}_1 \mid \mathbf{I}_t = \mathbf{x}, t)$ is therefore the sum of the block entropies: the uncertainty about a token is the sum of the uncertainties about its levels, position by position.

C.2 The posterior of the true vertex

Lemma C.2 (Log-odds representation). *In the single-block setting of Section 4, let $\mathbf{g} := \mathbf{x}_0/\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, so $\mathbf{I}_t = (1 - \alpha)\mathbf{e}_{i^*} + \alpha\sigma\mathbf{g}$. Then for every $j \neq i^*$,*

$$\log \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j \mid \mathbf{I}_t) - \log \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_{i^*} \mid \mathbf{I}_t) = \rho(g_j - g_{i^*}) - \rho^2, \quad \rho = \frac{1 - \alpha}{\alpha\sigma},$$

and consequently $\pi_* = \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_{i^*} \mid \mathbf{I}_t)$ is given by Eq. (11).

Proof. Idea. By Proposition 4.1 the posterior is a softmax of $\beta\mathbf{I}_t$, so in the ratio of two posterior probabilities the normaliser cancels and only $\beta(x_j - x_{i^*})$ remains. It suffices to insert the coordinates of $\mathbf{I}_t$ on $\{\mathbf{x}_1 = \mathbf{e}_{i^*}\}$ and to recognise ρ and ρ^2 in the result; summing the odds over j then gives $1/\pi_*$.

The single-block setting is Proposition 4.1 with $K = L = 1$, $v_1 = n$, uniform μ_1 , $\alpha = \alpha_1(t) \in (0, 1)$ and $\sigma = \sigma_1 > 0$; in particular $\rho > 0$. Since $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{g} = \mathbf{x}_0/\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and $\mathbf{I}_t = \alpha\mathbf{x}_0 + (1 - \alpha)\mathbf{x}_1 = (1 - \alpha)\mathbf{e}_{i^*} + \alpha\sigma\mathbf{g}$ on $\{\mathbf{x}_1 = \mathbf{e}_{i^*}\}$. By Eq. (10), with $\beta = (1 - \alpha)/(\alpha^2\sigma^2)$ and $\mathbf{x} = \mathbf{I}_t$, the normalising sum cancels in the ratio of two posterior probabilities, so

$$\log \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j \mid \mathbf{I}_t) - \log \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_{i^*} \mid \mathbf{I}_t) = \beta x_j - \beta x_{i^*} = \beta(x_j - x_{i^*}). \quad (60)$$

The coordinates of $\mathbf{x} = \mathbf{I}_t$ are $x_j = \alpha\sigma g_j$ for $j \neq i^*$ and $x_{i^*} = (1 - \alpha) + \alpha\sigma g_{i^*}$, because $(\mathbf{e}_{i^*})_j = \mathbb{1}\{j = i^*\}$. Hence, for $j \neq i^*$, using $\beta\alpha\sigma = (1 - \alpha)/(\alpha\sigma) = \rho$ and $\beta(1 - \alpha) = (1 - \alpha)^2/(\alpha^2\sigma^2) = \rho^2$,

$$\beta(x_j - x_{i^*}) = \beta\alpha\sigma(g_j - g_{i^*}) - \beta(1 - \alpha) = \frac{1 - \alpha}{\alpha\sigma}(g_j - g_{i^*}) - \frac{(1 - \alpha)^2}{\alpha^2\sigma^2} = \rho(g_j - g_{i^*}) - \rho^2.$$

For $j = i^*$ the left-hand side of Eq. (60) is 0. Since the posterior probabilities over $j \in [n]$ sum to one and $\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_{i^*} \mid \mathbf{I}_t) > 0$ (a softmax has positive entries),

$$\frac{1}{\pi_*} = \sum_{j=1}^n \frac{\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j \mid \mathbf{I}_t)}{\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_{i^*} \mid \mathbf{I}_t)} = 1 + \sum_{j \neq i^*} \exp(\rho(g_j - g_{i^*}) - \rho^2),$$

where the term $j = i^*$ equals 1 and the terms $j \neq i^*$ are the exponentials of the log-odds just computed. Taking reciprocals gives Eq. (11). $\square$

C.3 Gaussian tools

Let $g \sim \mathcal{N}(0, 1)$, $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$, $\Phi(x) = \mathbb{P}(g \leq x)$ and $Q(x) = 1 - \Phi(x) = \mathbb{P}(g > x)$. Since φ is even, $\Phi(-x) = Q(x)$ for all $x \in \mathbb{R}$; since the law of g has no atoms, $\mathbb{P}(g < x) = \Phi(x)$ and $\mathbb{P}(g \geq x) = Q(x)$.

Lemma C.3 (Upper tail). *For $x \geq 0$, $Q(x) \leq e^{-x^2/2}$.*

Proof. For $x = 0$ the claim is $Q(0) = \frac{1}{2} \leq 1$. Let $x > 0$ and $\lambda > 0$. Completing the square, $\lambda z - \frac{z^2}{2} = \frac{\lambda^2}{2} - \frac{(z - \lambda)^2}{2}$, so

$$\mathbb{E}e^{\lambda g} = \int_{\mathbb{R}} e^{\lambda z} \varphi(z) dz = e^{\lambda^2/2} \int_{\mathbb{R}} \varphi(z - \lambda) dz = e^{\lambda^2/2}.$$

As $y \mapsto e^{\lambda y}$ is increasing, $\{g > x\} \subset \{e^{\lambda g} \geq e^{\lambda x}\}$, and Markov's inequality for the nonnegative variable $e^{\lambda g}$ gives

$$Q(x) \leq \mathbb{P}(e^{\lambda g} \geq e^{\lambda x}) \leq e^{-\lambda x} \mathbb{E}e^{\lambda g} = e^{-\lambda x + \lambda^2/2}.$$

Taking $\lambda = x > 0$ gives $Q(x) \leq e^{-x^2 + x^2/2} = e^{-x^2/2}$. $\square$

Lemma C.4 (Maximum, upper tail). Let $g_1, \dots, g_m$ be independent standard normal, $m \geq 1$, $s = \sqrt{2 \log m}$ and $u \geq 0$. Then $\mathbb{P}(\max_i g_i > s + u) \leq e^{-u^2/2}$.

Proof. Since $m \geq 1$, $s \geq 0$ and $s + u \geq 0$. As $\{\max_i g_i > s + u\} = \bigcup_i \{g_i > s + u\}$, the union bound and Lemma C.3 (with $x = s + u \geq 0$) give

$$\mathbb{P}(\max_i g_i > s + u) \leq \sum_{i=1}^m \mathbb{P}(g_i > s + u) = m Q(s + u) \leq m e^{-(s+u)^2/2}.$$

Since $su \geq 0$, $(s + u)^2 = s^2 + 2su + u^2 \geq s^2 + u^2$, and $s^2/2 = \log m$, so $m e^{-(s+u)^2/2} \leq m e^{-s^2/2} e^{-u^2/2} = m e^{-\log m} e^{-u^2/2} = e^{-u^2/2}$. $\square$

Lemma C.5 (Truncated exponential moment). For $\rho \in \mathbb{R}$ and $M \in \mathbb{R}$, $\mathbb{E}[e^{\rho g} \mathbb{1}\{g \leq M\}] = e^{\rho^2/2} \Phi(M - \rho)$.

Proof. Completing the square, $\rho z - \frac{z^2}{2} = \frac{\rho^2}{2} - \frac{(z-\rho)^2}{2}$, so $e^{\rho z} \varphi(z) = e^{\rho^2/2} \varphi(z - \rho)$ for all z ; in particular the nonnegative integrand below is integrable. Therefore, with the substitution $y = z - \rho$,

$$\mathbb{E}[e^{\rho g} \mathbb{1}\{g \leq M\}] = \int_{-\infty}^M e^{\rho z} \varphi(z) dz = e^{\rho^2/2} \int_{-\infty}^M \varphi(z - \rho) dz = e^{\rho^2/2} \int_{-\infty}^{M-\rho} \varphi(y) dy = e^{\rho^2/2} \Phi(M - \rho). \quad \square$$

Lemma C.6 (Mills ratio, lower bound [20]). For $x > 0$, $Q(x) \geq \frac{x}{1+x^2} \varphi(x)$.

Proof. Let $r(x) = Q(x) - \frac{x}{1+x^2} \varphi(x)$ for $x \in \mathbb{R}$. We use $Q' = -\varphi$, $\varphi'(x) = -x\varphi(x)$ and $\frac{d}{dx} \frac{x}{1+x^2} = \frac{(1+x^2) - 2x^2}{(1+x^2)^2} = \frac{1-x^2}{(1+x^2)^2}$. By the product rule,

$$r'(x) = -\varphi(x) - \frac{1-x^2}{(1+x^2)^2} \varphi(x) + \frac{x^2}{1+x^2} \varphi(x) = -\varphi(x) \frac{(1+x^2)^2 + (1-x^2) - x^2(1+x^2)}{(1+x^2)^2}.$$

The numerator is $1 + 2x^2 + x^4 + 1 - x^2 - x^2 - x^4 = 2$, so $r'(x) = -2\varphi(x)/(1+x^2)^2 < 0$, and r is strictly decreasing. As $x \rightarrow \infty$, $Q(x) \rightarrow 0$ and $0 \leq \frac{x}{1+x^2} \varphi(x) \leq \varphi(x) \rightarrow 0$, so $r(x) \rightarrow 0$. Now fix $x > 0$. For every $y > x + 1$, $r(x + 1) > r(y)$; letting $y \rightarrow \infty$ gives $r(x + 1) \geq 0$, and hence $r(x) > r(x + 1) \geq 0$. $\square$

Lemma C.7 (Maximum, lower tail). With $g_1, \dots, g_m$ as in Lemma C.4 and $x \in \mathbb{R}$, $\mathbb{P}(\max_i g_i < x) \leq \exp(-mQ(x))$.

Proof. Since $\{\max_i g_i < x\} = \bigcap_i \{g_i < x\}$ and the g_i are independent with $\mathbb{P}(g_i < x) = \Phi(x) = 1 - Q(x)$,

$$\mathbb{P}(\max_i g_i < x) = \prod_{i=1}^m \mathbb{P}(g_i < x) = (1 - Q(x))^m.$$

Since $1 - y \leq e^{-y}$ for all real y (convexity of $y \mapsto e^{-y}$, whose tangent at 0 is $1 - y$), we have $0 \leq 1 - Q(x) \leq e^{-Q(x)}$, and raising these nonnegative numbers to the power m preserves the order: $(1 - Q(x))^m \leq e^{-mQ(x)}$. $\square$

C.4 Proof of Theorem 4.2

Proof. Idea. By Lemma C.2, $\pi_* = 1/(1 + \Sigma)$ with $\Sigma = e^{-\rho g_{i^*} - \rho^2} S$, where S is the sum of $e^{\rho g_j}$ over the m competitors, so both parts compare the size of S with the advantage e^{ρ^2} of the true vertex. For (a) we

need S to be small. Above the threshold the mean of S is carried by values $g_j \approx \rho$ that the maximum of the competitors does not reach, so we truncate at $M = s + u$, which no competitor exceeds except with probability $e^{-u^2/2}$ (Lemma C.4), and bound the truncated sum by Markov's inequality, its mean being explicit by Lemma C.5; the truncation improves the exponent from $(\rho^2 - s^2)/2$ to $\rho(\rho - s)$. For (b) we need S to be large, and one term suffices: by independence and the Mills ratio (Lemmas C.6 and C.7) the largest competitor exceeds $s - h$ with high probability, and this single term already outweighs e^{ρ^2} when $\rho < s - h - u$.

Throughout, $m = n - 1 \geq 1$, $s = s_n = \sqrt{2 \log m} \geq 0$, $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, so that $g_1, \dots, g_n$ are independent standard normal and the m variables $(g_j)_{j \neq i^*}$ satisfy the hypotheses of Lemmas C.4 and C.7. The signal-to-noise ratio $\rho = (1 - \alpha)/(\alpha\sigma)$ is a deterministic positive number. By Lemma C.2, factoring $e^{-\rho g_{i^*} - \rho^2}$ out of every term,

$$\pi_\star = \frac{1}{1 + \Sigma}, \quad \Sigma := \sum_{j \neq i^*} e^{\rho(g_j - g_{i^*}) - \rho^2} = e^{-\rho g_{i^*} - \rho^2} S, \quad S := \sum_{j \neq i^*} e^{\rho g_j}. \quad (61)$$

(a) Fix $u, u' > 0$ and $\rho \geq s + 2u$, and let $M := s + u$. Define the events

$$A = \{g_{i^*} \geq -u\}, \quad B = \{\max_{j \neq i^*} g_j \leq M\}, \quad C = \{S_M \leq e^{u'} \mathbb{E} S_M\}, \quad S_M := \sum_{j \neq i^*} e^{\rho g_j} \mathbb{1}_{\{g_j \leq M\}}.$$

Probabilities. $\mathbb{P}(A^c) = \mathbb{P}(g_{i^*} < -u) = \Phi(-u) = Q(u) \leq e^{-u^2/2}$ by symmetry and Lemma C.3 ($u > 0$). $B^c = \{\max_{j \neq i^*} g_j > s + u\}$, so $\mathbb{P}(B^c) \leq e^{-u^2/2}$ by Lemma C.4. By Lemma C.5 and linearity of expectation,

$$\mathbb{E} S_M = \sum_{j \neq i^*} \mathbb{E}[e^{\rho g_j} \mathbb{1}_{\{g_j \leq M\}}] = m e^{\rho^2/2} \Phi(M - \rho) \in (0, \infty). \quad (62)$$

Since $S_M \geq 0$ and $C^c = \{S_M > e^{u'} \mathbb{E} S_M\} \subset \{S_M \geq e^{u'} \mathbb{E} S_M\}$, Markov's inequality gives $\mathbb{P}(C^c) \leq \mathbb{E} S_M / (e^{u'} \mathbb{E} S_M) = e^{-u'}$. By the union bound,

$$\mathbb{P}(A \cap B \cap C) \geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) - \mathbb{P}(C^c) \geq 1 - 2e^{-u^2/2} - e^{-u'}.$$

The truncated sum. Because $\rho \geq s + 2u > s + u = M$, we have $\rho - M > 0$ and $\Phi(M - \rho) = Q(\rho - M) \leq e^{-(\rho - M)^2/2}$ by Lemma C.3. Inserting this into Eq. (62) and expanding $\frac{\rho^2}{2} - \frac{(\rho - M)^2}{2} = \frac{\rho^2 - \rho^2 + 2\rho M - M^2}{2} = \rho M - \frac{M^2}{2}$,

$$\mathbb{E} S_M \leq m \exp\left(\frac{\rho^2}{2} - \frac{(\rho - M)^2}{2}\right) = m \exp\left(\rho M - \frac{M^2}{2}\right). \quad (63)$$

The bound on the good event. Work on $A \cap B \cap C$. On B , every g_j with $j \neq i^*$ is at most M , so all indicators in S_M equal 1 and $S = S_M$. On C , $S_M \leq e^{u'} \mathbb{E} S_M$. On A , $-g_{i^*} \leq u$, so $-\rho g_{i^*} \leq \rho u$ as $\rho > 0$. Since $\Sigma \geq 0$, $1 - \pi_\star = \Sigma / (1 + \Sigma) \leq \Sigma$. Combining these with Eq. (61) and Eq. (63),

$$\begin{aligned} 1 - \pi_\star &\leq \Sigma = e^{-\rho g_{i^*} - \rho^2} S_M \leq e^{\rho u - \rho^2} e^{u'} \mathbb{E} S_M \leq e^{\rho u - \rho^2} e^{u'} m e^{\rho M - M^2/2} \\ &= \exp\left(u' + \rho u - \rho^2 + \rho M + \left(\log m - \frac{M^2}{2}\right)\right), \end{aligned}$$

using $m = e^{\log m}$ in the last step. Now $\log m = s^2/2$ and $M^2 = (s + u)^2 = s^2 + 2su + u^2$, so $\log m - M^2/2 = -su - u^2/2 \leq 0$. Dropping this nonpositive term and substituting $M = s + u$,

$$1 - \pi_\star \leq \exp(u' + \rho u - \rho^2 + \rho(s + u)) = \exp(u' - \rho(\rho - s - 2u)).$$

Thus $A \cap B \cap C \subset \{1 - \pi_\star \leq \exp(u' - \rho(\rho - s - 2u))\}$, and the probability bound for $A \cap B \cap C$ proves Eq. (12).

(b) Fix $u > 0$, $h \in (0, s)$ and $\eta > 0$ with $\rho \leq s - h - u - \eta$; in particular $s > h > 0$, so $\log m > 0$ and $m \geq 2$. Let $G := \max_{j \neq i^*} g_j$ and define $D = \{g_{i^*} \leq u\}$ and $E = \{G \geq s - h\}$.

Probabilities. $\mathbb{P}(D^c) = \mathbb{P}(g_{i^*} > u) = Q(u) \leq e^{-u^2/2}$ by Lemma C.3. Let $x := s - h > 0$. Then $E^c = \{G < x\}$; Lemma C.7 bounds its probability, and Lemma C.6 (as $x > 0$) together with the monotonicity of $y \mapsto e^{-y}$ gives

$$\mathbb{P}(E^c) = \mathbb{P}(G < x) \leq \exp(-mQ(x)) \leq \exp\left(-m \frac{x}{1 + x^2} \varphi(x)\right).$$

Since $(s-h)^2 = s^2 - 2sh + h^2$ and $me^{-s^2/2} = me^{-\log m} = 1$,

$$m\varphi(x) = \frac{m}{\sqrt{2\pi}} e^{-(s-h)^2/2} = \frac{m}{\sqrt{2\pi}} e^{-s^2/2} e^{sh-h^2/2} = \frac{1}{\sqrt{2\pi}} e^{sh-h^2/2}.$$

Hence $m \frac{x}{1+x^2} \varphi(x) = \frac{s-h}{1+(s-h)^2} \frac{e^{sh-h^2/2}}{\sqrt{2\pi}} = \Lambda_n(h)$ and $\mathbb{P}(E^c) \leq e^{-\Lambda_n(h)}$. By the union bound, $\mathbb{P}(D \cap E) \geq 1 - e^{-u^2/2} - e^{-\Lambda_n(h)}$.

The bound on the good event. The maximum G is attained at some $j_0 \neq i^*$. The sum Σ in Eq. (61) has nonnegative terms, one of which (the term $j = j_0$) is e^z with $z := \rho(G - g_{i^*}) - \rho^2$. Hence $\Sigma \geq e^z$ and, since $1 + e^z > e^z$,

$$\pi_\star = \frac{1}{1 + \Sigma} \leq \frac{1}{1 + e^z} \leq e^{-z}.$$

On $D \cap E$, $G - g_{i^*} \geq (s-h) - u$, so, as $\rho > 0$,

$$z \geq \rho(s-h-u) - \rho^2 = \rho(s-h-u-\rho) \geq \rho\eta,$$

where the last step multiplies $s-h-u-\rho \geq \eta$ (the hypothesis $\rho \leq s-h-u-\eta$) by $\rho > 0$. Hence $\pi_\star \leq e^{-z} \leq e^{-\rho\eta}$ on $D \cap E$, i.e. $D \cap E \subset \{\pi_\star \leq e^{-\rho\eta}\}$, which proves Eq. (13). $\square$

C.5 Proof of Corollary 4.3

Proof. Idea. Both statements follow from Theorem 4.2 once the free parameters are chosen as fixed fractions of $s = s_n$: the margins u, h, η then grow with n , while the hypotheses $\rho_n \geq s + 2u$ and $\rho_n \leq s - h - u - \eta$ still follow from $\rho_n \geq (1+\varepsilon)s$ and $\rho_n \leq (1-\varepsilon)s$. Every failure probability and every bound on the good event becomes a function of s alone, and it remains to check that each tends to zero as $s \rightarrow \infty$; the slowest is the term $1/s$ that comes from $u' = \log s$.

For each n , $\pi_\star$ is the random variable Eq. (11) with $\rho = \rho_n$ and $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and $0 < \pi_\star < 1$. As $n \rightarrow \infty$, $s = s_n = \sqrt{2 \log(n-1)} \rightarrow \infty$; for $n \geq 3$, $s \geq \sqrt{2 \log 2} > 1$. Fix $\delta \in (0, 1)$ (for $\delta \geq 1$ the events below are empty). Convergence $\pi_\star \rightarrow 1$ (resp. $\rightarrow 0$) in probability means $\mathbb{P}(1 - \pi_\star > \delta) \rightarrow 0$ (resp. $\mathbb{P}(\pi_\star > \delta) \rightarrow 0$) for every such δ .

Above the threshold. Let $\rho_n \geq (1+\varepsilon)s$ and $n \geq 3$. Apply Theorem 4.2(a) with $u = \varepsilon s/4 > 0$ and $u' = \log s > 0$. The hypothesis holds: $s + 2u = (1 + \varepsilon/2)s \leq (1 + \varepsilon)s \leq \rho_n$. Moreover $\rho_n - s - 2u \geq (1 + \varepsilon)s - (1 + \varepsilon/2)s = \varepsilon s/2 \geq 0$ and $\rho_n \geq (1 + \varepsilon)s \geq 0$, so the product of these two nonnegative quantities satisfies $\rho_n(\rho_n - s - 2u) \geq (1 + \varepsilon)s \cdot \varepsilon s/2$ and

$$b_n := u' - \rho_n(\rho_n - s - 2u) \leq \log s - \frac{\varepsilon(1+\varepsilon)}{2} s^2 \rightarrow -\infty,$$

because $\log s/s^2 \rightarrow 0$. Hence there is n_0 with $e^{b_n} \leq \delta$ for $n \geq n_0$, and for such n , $\{1 - \pi_\star \leq e^{b_n}\} \subset \{1 - \pi_\star \leq \delta\}$. By Eq. (12), with $u^2/2 = \varepsilon^2 s^2/32$ and $e^{-u'} = 1/s$,

$$\mathbb{P}(1 - \pi_\star > \delta) \leq \mathbb{P}(1 - \pi_\star > e^{b_n}) \leq 2e^{-\varepsilon^2 s^2/32} + \frac{1}{s} \rightarrow 0.$$

Below the threshold. Let $c \leq \rho_n \leq (1-\varepsilon)s$ (so $s > 0$, i.e. $n \geq 3$). Apply Theorem 4.2(b) with $h = u = \varepsilon s/4$ and $\eta = \varepsilon s/2$. The hypotheses hold: $u, \eta > 0$, $0 < h < s$ as $\varepsilon < 1$, and $s - h - u - \eta = s - \varepsilon s/4 - \varepsilon s/4 - \varepsilon s/2 = (1 - \varepsilon)s \geq \rho_n$. On the event of Eq. (13), $\pi_\star \leq e^{-\rho_n \eta} \leq e^{-c\varepsilon s/2}$, using $\rho_n \geq c$ and $\eta > 0$; as $e^{-c\varepsilon s/2} \rightarrow 0$, it is $\leq \delta$ for all large n , and then

$$\mathbb{P}(\pi_\star > \delta) \leq \mathbb{P}(\pi_\star > e^{-\rho_n \eta}) \leq e^{-u^2/2} + e^{-\Lambda_n(h)}.$$

First, $e^{-u^2/2} = e^{-\varepsilon^2 s^2/32} \rightarrow 0$. Second, let $x = s - h = (1 - \varepsilon/4)s$; then $x \leq s$, and $x \geq 1$ for $s \geq (1 - \varepsilon/4)^{-1}$, i.e. for all large n . For $x \geq 1$, $2x^2 \geq 1 + x^2$, i.e. $\frac{x}{1+x^2} \geq \frac{1}{2x} \geq \frac{1}{2s}$. Since $0 < \varepsilon < 1$ gives $\varepsilon^2 < \varepsilon$,

$$sh - \frac{h^2}{2} = \frac{\varepsilon s^2}{4} - \frac{\varepsilon^2 s^2}{32} \geq \frac{\varepsilon s^2}{4} - \frac{\varepsilon s^2}{32} = \frac{7\varepsilon s^2}{32} \geq \frac{\varepsilon s^2}{8}.$$

Therefore, by the definition of Λ_n in Eq. (13),

$$\Lambda_n(h) = \frac{x}{1+x^2} \frac{e^{sh-h^2/2}}{\sqrt{2\pi}} \geq \frac{e^{\varepsilon s^2/8}}{2\sqrt{2\pi}s} \rightarrow \infty,$$

since $e^{\varepsilon s^2/8}/s \rightarrow \infty$ as $s \rightarrow \infty$. Hence $e^{-\Lambda_n(h)} \rightarrow 0$, and $\mathbb{P}(\pi_\star > \delta) \rightarrow 0$. $\square$

C.6 Proof of Corollary 4.4

Proof. Idea. For the power schedule $\rho(t)$ is an explicit increasing function of t , so the threshold time is found by solving $\rho(t) = s_n$, and every statement of Theorem 4.2 about ρ becomes a statement about t by monotonicity. The proof computes the inverse, reads the dependence on n , σ and a off the closed form, and translates the two regimes.

Let $a > 0$ and $\sigma > 0$. With $\alpha(t) = (1-t)^a \in (0, 1]$ for $t \in [0, 1)$,

$$\rho(t) = \frac{1 - \alpha(t)}{\alpha(t)\sigma} = \frac{\alpha(t)^{-1} - 1}{\sigma} = \frac{(1-t)^{-a} - 1}{\sigma}.$$

By the chain rule, $\frac{d}{dt}(1-t)^{-a} = -a(1-t)^{-a-1} \cdot (-1)$, so $\rho'(t) = a(1-t)^{-a-1}/\sigma > 0$ on $[0, 1)$. Moreover $\rho(0) = 0$ and $\rho(t) \rightarrow \infty$ as $t \rightarrow 1$, since $(1-t)^{-a} \rightarrow \infty$. A continuous strictly increasing function on $[0, 1)$ with these boundary values is, by the intermediate value theorem, a bijection $[0, 1) \rightarrow [0, \infty)$. For $\rho \geq 0$, solving $\rho = ((1-t)^{-a} - 1)/\sigma$ gives $(1-t)^{-a} = 1 + \sigma\rho$, i.e. $1-t = (1 + \sigma\rho)^{-1/a}$, so $t(\rho) = 1 - (1 + \sigma\rho)^{-1/a}$. With $\rho = s_n = \sqrt{2\log(n-1)} \geq 0$ this is $t^\star = t(s_n)$, i.e. Eq. (14).

Monotonicity. Write $t^\star = 1 - \exp(-y)$ with $y := \ln(1 + \sigma s_n)/a \geq 0$; the map $y \mapsto 1 - e^{-y}$ is strictly increasing. The map $n \mapsto s_n$ is strictly increasing on $n \geq 2$, and $\ln(1 + \sigma s_n)$ is nondecreasing in σ and in s_n (strictly when the other factor of σs_n is positive). Hence y , and with it $t^\star$, is increasing in n and in σ ; and since $y = \ln(1 + \sigma s_n) \cdot a^{-1}$ with $\ln(1 + \sigma s_n) \geq 0$, y is nonincreasing in a , strictly when $\sigma s_n > 0$, i.e. $n \geq 3$. (For $n = 2$, $s_2 = 0$ and $t^\star = 0$ for all a, σ .)

Time windows. Fix $t \in (0, 1)$, so $\alpha(t) \in (0, 1)$ and Theorem 4.2 applies with $\rho = \rho(t)$. Because $\rho(\cdot)$ is strictly increasing with inverse $t(\cdot)$, for every $\rho_0 \geq 0$ we have $t \geq t(\rho_0) \iff \rho(t) \geq \rho_0$ and $t \leq t(\rho_0) \iff \rho(t) \leq \rho_0$. Let $u, u', \delta > 0$ and $t \geq t(s_n + 2u + \delta)$. Then $\rho(t) \geq s_n + 2u + \delta$, so $\rho(t) - s_n - 2u \geq \delta$ and $\rho(t)(\rho(t) - s_n - 2u) \geq (s_n + 2u + \delta)\delta$, and Eq. (12) gives

$$\mathbb{P}\left(1 - \pi_\star \leq \exp(u' - (s_n + 2u + \delta)\delta)\right) \geq 1 - 2e^{-u^2/2} - e^{-u'}.$$

Let $u, \eta > 0$, $h \in (0, s_n)$ with $s_n - h - u - \eta \geq 0$, and $t \in (0, t(s_n - h - u - \eta)]$. Then $\rho(t) \leq s_n - h - u - \eta$, and Eq. (13) gives $\mathbb{P}(\pi_\star \leq e^{-\rho(t)\eta}) \geq 1 - e^{-u^2/2} - e^{-\Lambda_n(h)}$. These are the two stated regimes. $\square$

C.7 Proof of Theorem 4.5

Proof. Idea. We must reduce each block to the single-block problem of Theorem 4.2. Two facts are needed: the posterior of a block depends on $\mathbf{I}_t$ only through that block, which is Proposition 4.1, and the noise of the block is a standard Gaussian vector independent of its true symbol, so that $\pi_{\star}^{\ell,k}(t)$ has exactly the law Eq. (11) with $n = v_k$ and $\rho = \rho_k(t)$. Steps 1–3 establish this; Step 4 applies Corollary 4.4 level by level with a union bound over positions, and Step 5 compares the closed-form thresholds.

Throughout, $v_k \geq 2$ for all k (implicit in Eq. (15)), and we write $s_v := \sqrt{2\log(v-1)}$.

Step 1: conditional independence structure. Fix $t \in (0, 1)$. Since t is independent of $(\mathbf{x}_0, \mathbf{x}_1)$, conditioning on t does not change the joint law of $(\mathbf{x}_0, \mathbf{x}_1)$. Under $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$ the blocks $\mathbf{x}_1^{\ell,k}$ are independent and uniform on $\{\mathbf{e}_1, \dots, \mathbf{e}_{v_k}\}$. The covariance Σ is diagonal, so the coordinates of the Gaussian vector $\mathbf{x}_0$ are independent, and the blocks $\mathbf{g}^{\ell,k} := \mathbf{x}_0^{\ell,k}/\sigma_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{v_k})$ are independent. As $\mathbf{x}_0$ and $\mathbf{x}_1$ are independent, the joint law of the family $((\mathbf{x}_1^{\ell,k}, \mathbf{g}^{\ell,k}))_{\ell,k}$ is the product of the laws of its members, i.e. the pairs $(\mathbf{x}_1^{\ell,k}, \mathbf{g}^{\ell,k})$ are independent across blocks, and within each pair $\mathbf{x}_1^{\ell,k}$ and $\mathbf{g}^{\ell,k}$ are independent. By Eq. (3),

$$\mathbf{I}_t^{\ell,k} = (1 - \alpha_k) \mathbf{e}_{i_{\ell,k}^\star} + \alpha_k \sigma_k \mathbf{g}^{\ell,k}, \quad \alpha_k = \alpha_k(t),$$

where $i_{\ell,k}^* \in [v_k]$ is the random index with $\mathbf{x}_1^{\ell,k} = \mathbf{e}_{i_{\ell,k}^*}$.

Step 2: the block posterior. By [Proposition 4.1](#), the posterior of $\mathbf{x}_1$ given $\mathbf{I}_t = \mathbf{x}$ is the product $\prod_{\ell',k'} q^{\ell',k'}(j_{\ell',k'} | \mathbf{x}^{\ell',k'})$ of the block posteriors [Eq. \(10\)](#). Summing over all configurations with $j_{\ell,k} = i$ and using [Eq. \(57\)](#) for the other blocks, each of whose factors sums to one,

$$\mathbb{P}(\mathbf{x}_1^{\ell,k} = \mathbf{e}_i | \mathbf{I}_t = \mathbf{x}, t) = q^{\ell,k}(i | \mathbf{x}^{\ell,k}) = \mathbb{P}(\mathbf{x}_1^{\ell,k} = \mathbf{e}_i | \mathbf{I}_t^{\ell,k} = \mathbf{x}^{\ell,k}, t).$$

So $\pi_{\star}^{\ell,k}(t) = q^{\ell,k}(i_{\ell,k}^* | \mathbf{I}_t^{\ell,k})$, whether one conditions on $\mathbf{I}_t$ or on $\mathbf{I}_t^{\ell,k}$. Define, for $n \geq 2$, $\rho > 0$, $i \in [n]$ and $\mathbf{y} \in \mathbb{R}^n$,

$$\Psi_{n,\rho}(i, \mathbf{y}) := \left(1 + \sum_{j \neq i} \exp(\rho(y_j - y_i) - \rho^2)\right)^{-1}. \quad (64)$$

The block posterior is the single-block softmax with $n = v_k$, $\alpha = \alpha_k$, $\sigma = \sigma_k$, so the computation in the proof of [Lemma C.2](#) (which uses only the representation of $\mathbf{I}_t^{\ell,k}$ in Step 1) gives

$$\pi_{\star}^{\ell,k}(t) = \Psi_{v_k, \rho_k(t)}(i_{\ell,k}^*, \mathbf{g}^{\ell,k}), \quad \rho_k(t) = \frac{1 - \alpha_k(t)}{\alpha_k(t) \sigma_k}. \quad (65)$$

Step 3: the law. Since $\mathbf{g}^{\ell,k}$ is independent of $i_{\ell,k}^*$, conditionally on $\{i_{\ell,k}^* = i\}$ the vector $\mathbf{g}^{\ell,k}$ is still $\mathcal{N}(\mathbf{0}, \mathbf{I}_{v_k})$, and $\pi_{\star}^{\ell,k}(t)$ has the law of $\Psi_{v_k, \rho_k(t)}(i, \mathbf{g})$ with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{v_k})$. This law does not depend on i : if P is the coordinate permutation exchanging i and i' , then $(P\mathbf{y})_{i'} = y_i$ and $\{(P\mathbf{y})_j : j \neq i'\} = \{y_j : j \neq i\}$ as multisets, so $\Psi(i', P\mathbf{y}) = \Psi(i, \mathbf{y})$, while $P\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{v_k})$. Averaging over i , $\pi_{\star}^{\ell,k}(t)$ has the law of [Eq. \(11\)](#) with $n = v_k$ and $\rho = \rho_k(t)$. By [Eq. \(65\)](#), $\pi_{\star}^{\ell,k}(t)$ is a measurable function of the pair $(\mathbf{x}_1^{\ell,k}, \mathbf{g}^{\ell,k})$; as these pairs are independent across blocks (Step 1), so are the $\pi_{\star}^{\ell,k}(t)$.

Step 4: thresholds and the union bound. [Theorem 4.2](#) is a statement about the law of [Eq. \(11\)](#), so it applies to each $\pi_{\star}^{\ell,k}(t)$ with $n = v_k$ and $\rho = \rho_k(t)$. Since $\alpha_k(t) = (1 - t)^{a_k}$, [Corollary 4.4](#) with $(a, \sigma, n) = (a_k, \sigma_k, v_k)$ shows that $\rho_k(\cdot)$ crosses s_{v_k} at $t_k^* = 1 - (1 + \sigma_k s_{v_k})^{-1/a_k}$, which is [Eq. \(15\)](#). For the simultaneous statement, fix $k, u, u' > 0$ and t with $\rho_k(t) \geq s_{v_k} + 2u$ (the same condition for every position ℓ), and let E_{ℓ} be the event in [Eq. \(12\)](#) for block (ℓ, k) . Then $\mathbb{P}(E_{\ell}^c) \leq 2e^{-u^2/2} + e^{-u'}$ for each ℓ , and by the union bound

$$\mathbb{P}\left(\bigcap_{\ell=1}^L E_{\ell}\right) \geq 1 - \sum_{\ell=1}^L \mathbb{P}(E_{\ell}^c) \geq 1 - L(2e^{-u^2/2} + e^{-u'})$$

(no independence is needed here).

Step 5: the ordering. Let $X_k := 1 + \sigma_k s_{v_k} \geq 1$, so $\ln X_k \geq 0$ and $t_k^* = 1 - X_k^{-1/a_k} = 1 - \exp(-\ln X_k/a_k)$. Since $y \mapsto 1 - e^{-y}$ is strictly increasing,

$$t_k^* < t_{k'}^* \iff \frac{\ln X_k}{a_k} < \frac{\ln X_{k'}}{a_{k'}}.$$

Let $k < k'$. Since $\sigma_k s_{v_k}$ is nondecreasing in k , $X_k \leq X_{k'}$ and $\ln X_k \leq \ln X_{k'}$; since $a_k > a_{k'} > 0$, $1/a_k < 1/a_{k'}$. Hence

$$\frac{\ln X_k}{a_k} \leq \frac{\ln X_{k'}}{a_k} < \frac{\ln X_{k'}}{a_{k'}},$$

where the first step divides $\ln X_k \leq \ln X_{k'}$ by $a_k > 0$ and the strict inequality multiplies $1/a_k < 1/a_{k'}$ by $\ln X_{k'} > 0$. The latter holds because $k' \geq 2$ gives $v_{k'} \geq 3$, so $s_{v_{k'}} \geq \sqrt{2 \log 2} > 0$ and $\sigma_{k'} s_{v_{k'}} > 0$. (For a binary level, $s_2 = \sqrt{2 \log 1} = 0$ and the threshold is $t^* = 0$: two vertices are distinguished as soon as ρ exceeds the margin $2u$.) This proves $t_1^* < \dots < t_K^*$ when $v_k \geq 3$ for $k \geq 2$. $\square$

C.8 Proof of [Corollary 4.6](#)

Proof. With $a_k = 1$, [Eq. \(15\)](#) reads

$$t_k^* = 1 - (1 + \sigma_k s_k)^{-1} = \frac{(1 + \sigma_k s_k) - 1}{1 + \sigma_k s_k} = \frac{\sigma_k s_k}{1 + \sigma_k s_k} = f(\sigma_k s_k), \quad f(y) := \frac{y}{1 + y} = 1 - \frac{1}{1 + y}.$$

Since $f'(y) = (1 + y)^{-2} > 0$, f is strictly increasing on $[0, \infty)$, and $\sigma_k s_k \geq 0$. Hence, for $k < k'$, $\sigma_k s_k < \sigma_{k'} s_{k'}$ implies $t_k^* = f(\sigma_k s_k) < f(\sigma_{k'} s_{k'}) = t_{k'}^*$. ([Theorem 4.5](#) is not needed here: its ordering hypothesis $a_1 > \dots > a_K$ fails for equal exponents, and the ordering follows directly from [Eq. \(15\)](#).) $\square$

C.9 Proof of Proposition 4.24

Proof. Idea. Everything reduces to the signal-to-noise ratio. If scaling the vertices by c multiplies ρ_k by c and leaves the posterior otherwise of the form Eq. (11), then Theorem 4.2 gives both thresholds directly and the comparison becomes the elementary inequality $(v-1)^K \leq v^K - 1$. The scaling claim is proved by dividing the interpolant by c , which turns it into the unit-vertex interpolant with prior scales σ_k/c .

Signal-to-noise ratio under scaling. With scaled vertices ce_i , the data are $\mathbf{x}'_1 = c\mathbf{x}_1$ with $\mathbf{x}_1$ as in Assumption 3.2, and the interpolant is $\mathbf{I}'_t = \mathbf{A}(t)\mathbf{x}_0 + (\mathbf{I} - \mathbf{A}(t))c\mathbf{x}_1$; its block (ℓ, k) is $\mathbf{I}'^{\ell, k}_t = (1 - \alpha_k)c\mathbf{e}_{i^*} + \alpha_k\sigma_k\mathbf{g}$, with $\mathbf{g} = \mathbf{g}^{\ell, k}$ as in the proof of Theorem 4.5. Dividing by $c > 0$,

$$\mathbf{I}'_t/c = \mathbf{A}(t)(\mathbf{x}_0/c) + (\mathbf{I} - \mathbf{A}(t))\mathbf{x}_1, \quad \mathbf{x}_0/c \sim \mathcal{N}(\mathbf{0}, \text{diag}((\sigma_k/c)^2 \mathbf{I}_{v_k})_{\ell, k}),$$

which is the interpolant of Assumption 3.2 with unit vertices and scales σ_k/c . Since $\mathbf{x} \mapsto \mathbf{x}/c$ is a bijection of $\mathbb{R}^D$, conditioning on $\mathbf{I}'_t$ is the same as conditioning on $\mathbf{I}'_t/c$ (in Bayes' rule with densities, the Jacobian factor c^{-D} is common to all configurations and cancels). Hence Proposition 4.1 applies: the posterior factorises over blocks, and the block posterior given $\mathbf{I}'^{\ell, k}_t = \mathbf{x}$ is Eq. (10) evaluated at $\mathbf{x}/c$ with scale σ_k/c , i.e. a softmax in $\mathbf{x}$ with inverse temperature

$$\beta' = \frac{1 - \alpha_k}{\alpha_k^2(\sigma_k/c)^2} \cdot \frac{1}{c} = \frac{c(1 - \alpha_k)}{\alpha_k^2\sigma_k^2}.$$

Equivalently, the likelihood of vertex i is proportional to $\exp(-\|\mathbf{x} - (1 - \alpha_k)c\mathbf{e}_i\|^2 / (2\alpha_k^2\sigma_k^2))$, and $\|\mathbf{x} - (1 - \alpha_k)c\mathbf{e}_i\|^2 = \|\mathbf{x}\|^2 - 2c(1 - \alpha_k)x_i + c^2(1 - \alpha_k)^2$, where only the middle term depends on i . The coordinates of $\mathbf{x} = \mathbf{I}'^{\ell, k}_t$ are $x_j = \alpha_k\sigma_k g_j$ for $j \neq i^*$ and $x_{i^*} = c(1 - \alpha_k) + \alpha_k\sigma_k g_{i^*}$, so as in Lemma C.2, for $j \neq i^*$ the log-odds of j against i^* are

$$\beta'(x_j - x_{i^*}) = \beta'(\alpha_k\sigma_k(g_j - g_{i^*}) - c(1 - \alpha_k)) = \frac{c(1 - \alpha_k)}{\alpha_k\sigma_k}(g_j - g_{i^*}) - \frac{c^2(1 - \alpha_k)^2}{\alpha_k^2\sigma_k^2} = c\rho_k(g_j - g_{i^*}) - (c\rho_k)^2.$$

Summing the exponentials as in the proof of Lemma C.2 and arguing as in Step 3 of the proof of Theorem 4.5, $\pi_{\star}^{\ell, k}$ has the law of Eq. (11) with ρ replaced by $c\rho_k$ (which is also the signal-to-noise ratio $(1 - \alpha_k)/(\alpha_k\sigma_k/c)$ of the rescaled model).

Thresholds. With $v_k = v$, $a_k = a$ and $\sigma_k = \sigma$ for all k , all levels have the same $\rho_k(t) = \rho(t) = ((1 - t)^{-a} - 1)/\sigma$. With $c = K^{-1/2}$ every block, in particular the finest one, is a single block with $n = v$ and signal-to-noise ratio $\rho/\sqrt{K}$, so by Theorem 4.2 its threshold is $\rho/\sqrt{K} = s_v = \sqrt{2\log(v-1)}$, i.e. $\rho = \sqrt{K}s_v = \sqrt{2K\log(v-1)}$. A flat token is a single block ($K = 1$, $c = 1$) with $n = V$ vertices, the same schedule and scale, hence signal-to-noise ratio ρ and threshold $\rho = s_V = \sqrt{2\log(V-1)}$.

Inequality. Let $K \geq 1$. By the binomial theorem, keeping only the terms $i = K$ and $i = 0$ (distinct as $K \geq 1$) of a sum of nonnegative terms,

$$V = v^K = ((v-1) + 1)^K = \sum_{i=0}^K \binom{K}{i} (v-1)^i \geq (v-1)^K + 1,$$

so $1 \leq (v-1)^K \leq V-1$ (as $v \geq 2$). Since $\log$ is increasing, $K\log(v-1) = \log(v-1)^K \leq \log(V-1)$; both sides are ≥ 0 , and multiplying by 2 and taking square roots (increasing on $[0, \infty)$) gives the first part of Eq. (24).

Limit. Fix $K \geq 1$ and let $v \rightarrow \infty$ (for $v \geq 3$ all logarithms below are positive). Since $v-1 = v(1-1/v)$ and $v^K - 1 = v^K(1-v^{-K})$,

$$K\log(v-1) = K\log v + K\log(1-1/v), \quad \log(v^K - 1) = K\log v + \log(1-v^{-K}),$$

and dividing numerator and denominator by $K\log v > 0$,

$$\frac{K\log(v-1)}{\log(v^K - 1)} = \frac{1 + \log(1-1/v)/\log v}{1 + \log(1-v^{-K})/(K\log v)}.$$

As $v \rightarrow \infty$, $\log(1-1/v) \rightarrow 0$, $\log(1-v^{-K}) \rightarrow 0$ and $\log v \rightarrow \infty$, so both correction terms tend to 0; the numerator and the denominator tend to 1, and the ratio tends to 1. $\square$

C.10 Proofs for Section 4.4

The proofs build on one another as follows. [Theorem 4.7](#) rescales time so that every block relaxes linearly towards its target, which gives the prior noise remaining in each block exactly, for any data. [Theorem 4.9](#) uses the symmetry of product-uniform data to fix the vertex that each block approaches, and [Theorem 4.10](#) combines the two results into lower and upper bounds on arrival times. The upper bound requires the integral $K_{\ell,k}$ to be finite, which [Proposition 4.12](#) derives from the local analysis near a vertex; [Corollary 4.13](#) then compares the bounds of two levels as $\varepsilon \rightarrow 0$.

Throughout, [Assumption 3.2](#) holds, $\mathcal{O} := \mathbb{R}^D \setminus \mathcal{M}_1$, and w_j, q are as in [Eq. \(4\)](#). Since $\mathcal{M}_1^+ \subset \mathcal{M}_1$, we have $\mathcal{O} \subset \Omega$, so [Lemma 3.6](#) applies on $\mathcal{O}$. If $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$, then $\pi_j = \prod_{\ell,k} v_k^{-1} > 0$ for every j , hence $\mathcal{M}_1^+ = \mathcal{M}_1$ and $\Omega = \mathcal{O}$. We use the conventions $\ln 0 := -\infty$, $\ln \infty := \infty$ and $\inf \emptyset := \infty$.

Lemma C.8 (Continuity of the speed and the targets). *M and $\bar{p}^{\ell,k}$ are continuous on Ω .*

Proof. Idea. Both functions are ratios of time integrals of the likelihood, so it suffices that these integrals depend continuously on $\mathbf{x}$ and that the denominators are positive. Continuity follows from dominated convergence, with the local bound of [Lemma B.3\(iv\)](#) as the dominating function.

With Z_j, Q_j from [Eq. \(43\)](#), $M = \sum_j Z_j / \sum_j Q_j$ and $\bar{p}^{\ell,k} = \sum_j Z_j e_{j\ell,k} / \sum_j Z_j$. Fix $\bar{\mathbf{x}} \in \Omega$ and let U and G_0 be as in [Lemma B.3\(iv\)](#), in the version on Ω established in the proof of [Lemma 3.6](#). For $\mathbf{x} \in U$, $t \in (0, 1)$ and every j , taking $\beta = \mathbf{0}$ and $k = 1$ there,

$$p(\mathbf{x} \mid t, j) \leq G_0(t), \quad \frac{p(\mathbf{x} \mid t, j)}{1-t} \leq \frac{G_0(t)}{a_1},$$

and $\mathbf{x} \mapsto p(\mathbf{x} \mid t, j)$ is continuous. By dominated convergence, Z_j and Q_j are continuous on U . The denominators are positive: $\sum_j Q_j = q > 0$ ([Lemma B.3\(ii\)](#)) and $\sum_j Z_j = Mq \geq q > 0$ ([Lemma 3.6](#)). Hence M and $\bar{p}^{\ell,k}$ are continuous on U , and $\bar{\mathbf{x}}$ was arbitrary. $\square$

Lemma C.9 (Dini derivatives under composition). *Let $I, J \subset \mathbb{R}$ be intervals, $g : J \rightarrow \mathbb{R}$ continuous, and $\varsigma_0 \in I$ not the right endpoint of I .*

- (a) *If $\sigma : I \rightarrow J$ is continuous, strictly increasing and differentiable at ς_0 with $\sigma'(\varsigma_0) > 0$, and $D^+g(\sigma(\varsigma_0)) < \infty$, then $D^+(g \circ \sigma)(\varsigma_0) = \sigma'(\varsigma_0) D^+g(\sigma(\varsigma_0))$.*
- (b) *If $G : I \rightarrow \mathbb{R}$ is continuous with $D^+G(\varsigma_0) < \infty$, and ϕ is C^1 on an open interval containing $G(I)$ with $\phi' > 0$, then $D^+(\phi \circ G)(\varsigma_0) = \phi'(G(\varsigma_0)) D^+G(\varsigma_0)$.*
- (c) *If $G : I \rightarrow \mathbb{R}$ and $H : I \rightarrow \mathbb{R}$ is differentiable at ς_0 , then $D^+(G + H)(\varsigma_0) = D^+G(\varsigma_0) + H'(\varsigma_0)$.*

Here $c \cdot (-\infty) := -\infty$ for $c > 0$.

Proof. We use: if $b_h \rightarrow b > 0$ and $\limsup_{h \downarrow 0} a_h = A \in [-\infty, \infty)$, then $\limsup_{h \downarrow 0} a_h b_h = bA$; and $\limsup(a_h + b_h) = \limsup a_h + \lim b_h$ when $\lim b_h$ exists in $\mathbb{R}$. (a) For small $h > 0$ put $k(h) := \sigma(\varsigma_0 + h) - \sigma(\varsigma_0) > 0$. Then

$$\frac{g(\sigma(\varsigma_0 + h)) - g(\sigma(\varsigma_0))}{h} = \frac{g(\sigma(\varsigma_0) + k(h)) - g(\sigma(\varsigma_0))}{k(h)} \cdot \frac{k(h)}{h}.$$

The map $h \mapsto k(h)$ is a continuous strictly increasing bijection of some $(0, h_0]$ onto $(0, k(h_0)]$, so the $\limsup_{h \downarrow 0}$ of the first factor equals $D^+g(\sigma(\varsigma_0))$, and the second factor tends to $\sigma'(\varsigma_0) > 0$. (b) By the mean value theorem, $\phi(G(\varsigma_0 + h)) - \phi(G(\varsigma_0)) = \phi'(\xi_h)(G(\varsigma_0 + h) - G(\varsigma_0))$ with ξ_h between $G(\varsigma_0)$ and $G(\varsigma_0 + h)$; by continuity of G and ϕ' , $\phi'(\xi_h) \rightarrow \phi'(G(\varsigma_0)) > 0$. (c) The difference quotient of H converges to $H'(\varsigma_0)$. $\square$

Proof of Theorem 4.7. *Idea.* We cannot integrate the flow, but by [Lemma 3.6](#) each block moves towards a point of its own simplex at the rate $a_k M$, with one scalar M for all blocks. Measuring time in units of $1/M$ removes this common factor and leaves for each block the linear equation $d\mathbf{Z}^{\ell,k}/d\varsigma = a_k(\mathbf{f} - \mathbf{Z}^{\ell,k})$ with a forcing $\mathbf{f}$ in the simplex, which variation of constants solves exactly (Steps 1–3); this is [Eq. \(16\)](#). It remains

to show that the flow time does not stop at a finite value: the hull coordinate decays only like $e^{-a_k \varsigma}$ by Eq. (70), which keeps the trajectory a positive distance from $\mathcal{M}_1$ on every bounded flow-time interval (Steps 5 and 6).

Step 1: existence and the flow time. $\mathbf{b}$ is C^∞ on the open set $\mathcal{O}$ (Proposition 3.3), hence locally Lipschitz there (mean value inequality on closed balls in $\mathcal{O}$). By the Picard–Lindelöf theorem and the standard continuation argument there is a unique maximal forward solution $\mathbf{X} : [0, S) \rightarrow \mathcal{O}$, $S \in (0, \infty]$, with $\mathbf{X}_0 = \mathbf{x}_0$; it is C^1 because $\mathbf{b}$ is continuous. By Lemma C.8 and Lemma 3.6, the function $r \mapsto M(\mathbf{X}_r)$ is continuous on $[0, S)$ with $M(\mathbf{X}_r) \geq 1$. By the fundamental theorem of calculus, $\varsigma(s) = \int_0^s M(\mathbf{X}_r) dr$ is C^1 with

$$\varsigma'(s) = M(\mathbf{X}_s) \geq 1 \quad (s \in [0, S)). \quad (66)$$

Hence ς is continuous and strictly increasing with $\varsigma(0) = 0$; the limit $\varsigma_{\max} := \lim_{s \uparrow S} \varsigma(s) \in (0, \infty]$ exists by monotonicity, and by the intermediate value theorem ς is a bijection of $[0, S)$ onto $[0, \varsigma_{\max})$. By the inverse function theorem in one variable (the derivative Eq. (66) is continuous and nonzero), the inverse $s(\cdot)$ is C^1 with

$$s'(\varsigma) = \frac{1}{\varsigma'(s(\varsigma))} = \frac{1}{M(\mathbf{X}_{s(\varsigma)})} \quad (\varsigma \in [0, \varsigma_{\max})). \quad (67)$$

Step 2: the equation in flow time. $\mathbf{Z}(\varsigma) = \mathbf{X}_{s(\varsigma)} \in \mathcal{O}$ is C^1 as a composition of C^1 maps. By the chain rule, Eq. (67) and Eq. (7) (valid on $\mathcal{O} \subset \Omega$),

$$\frac{d}{d\varsigma} \mathbf{Z}^{\ell,k}(\varsigma) = \dot{\mathbf{X}}_{s(\varsigma)}^{\ell,k} s'(\varsigma) = \frac{\mathbf{b}^{\ell,k}(\mathbf{Z}(\varsigma))}{M(\mathbf{Z}(\varsigma))} = a_k(\bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma)) - \mathbf{Z}^{\ell,k}(\varsigma)).$$

Step 3: variation of constants. Fix a block and write $\mathbf{f}(\varsigma) := \bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma))$. It is continuous on $[0, \varsigma_{\max})$ (Lemma C.8) with values in Δ^{v_k-1} (Lemma 3.6). By the product rule and Step 2,

$$\frac{d}{d\varsigma} \left(e^{a_k \varsigma} \mathbf{Z}^{\ell,k}(\varsigma) \right) = e^{a_k \varsigma} \left(a_k \mathbf{Z}^{\ell,k} + \frac{d\mathbf{Z}^{\ell,k}}{d\varsigma} \right) = e^{a_k \varsigma} \left(a_k \mathbf{Z}^{\ell,k} + a_k \mathbf{f} - a_k \mathbf{Z}^{\ell,k} \right) = a_k e^{a_k \varsigma} \mathbf{f}(\varsigma).$$

The right-hand side is continuous, so the fundamental theorem of calculus on $[0, \varsigma]$ and $\mathbf{Z}^{\ell,k}(0) = \mathbf{X}_{s(0)}^{\ell,k} = \mathbf{X}_0^{\ell,k} = \mathbf{x}_0^{\ell,k}$ give $e^{a_k \varsigma} \mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{x}_0^{\ell,k} = \int_0^\varsigma a_k e^{a_k \varsigma'} \mathbf{f}(\varsigma') d\varsigma'$; multiplying by $e^{-a_k \varsigma}$,

$$\mathbf{Z}^{\ell,k}(\varsigma) = e^{-a_k \varsigma} \mathbf{x}_0^{\ell,k} + \int_0^\varsigma a_k e^{-a_k(\varsigma - \varsigma')} \mathbf{f}(\varsigma') d\varsigma'. \quad (68)$$

For fixed ς the kernel $\kappa_\varsigma(\varsigma') := a_k e^{-a_k(\varsigma - \varsigma')}$ is positive on $[0, \varsigma]$ with

$$\int_0^\varsigma \kappa_\varsigma(\varsigma') d\varsigma' = \left[e^{-a_k(\varsigma - \varsigma')} \right]_{\varsigma'=0}^{\varsigma'=\varsigma} = 1 - e^{-a_k \varsigma} = 1 - \lambda_k(\varsigma). \quad (69)$$

For $\varsigma > 0$ we have $1 - \lambda_k(\varsigma) \in (0, 1)$ and set $\mathbf{y}^{\ell,k}(\varsigma) := (1 - \lambda_k(\varsigma))^{-1} \int_0^\varsigma \kappa_\varsigma \mathbf{f}$; for $\varsigma = 0$ set $\mathbf{y}^{\ell,k}(0) := \mathbf{f}(0) = \bar{\mathbf{p}}^{\ell,k}(\mathbf{x}_0)$. Coordinatewise $y_i^{\ell,k} = (1 - \lambda_k)^{-1} \int \kappa_\varsigma f_i \geq 0$, and by $\mathbf{1}^\top \mathbf{f} = 1$ and Eq. (69), $\sum_i y_i^{\ell,k} = (1 - \lambda_k)^{-1} \int \kappa_\varsigma = 1$; so $\mathbf{y}^{\ell,k}(\varsigma) \in \Delta^{v_k-1}$. For $\varsigma > 0$, Eq. (68) reads $\mathbf{Z}^{\ell,k} = \lambda_k \mathbf{x}_0^{\ell,k} + (1 - \lambda_k) \mathbf{y}^{\ell,k}$, which is Eq. (16); at $\varsigma = 0$ both sides of Eq. (16) equal $\mathbf{x}_0^{\ell,k}$ because $\lambda_k(0) = 1$. For $\varsigma > 0$, Eq. (16) determines $\mathbf{y}^{\ell,k}(\varsigma) = (\mathbf{Z}^{\ell,k}(\varsigma) - \lambda_k \mathbf{x}_0^{\ell,k}) / (1 - \lambda_k)$ uniquely, and $\mathbf{y}^{\ell,k}$ is continuous on $[0, \varsigma_{\max})$: it is continuous for $\varsigma > 0$ by this formula, and $\|\mathbf{y}^{\ell,k}(\varsigma) - \mathbf{f}(0)\| \leq (1 - \lambda_k)^{-1} \int_0^\varsigma \kappa_\varsigma \|\mathbf{f} - \mathbf{f}(0)\| \leq \max_{[0, \varsigma]} \|\mathbf{f} - \mathbf{f}(0)\| \rightarrow 0$ as $\varsigma \downarrow 0$. Finally, with $t := 1 - e^{-\varsigma} \in [0, 1)$, $\lambda_k(\varsigma) = e^{-a_k \varsigma} = (e^{-\varsigma})^{a_k} = (1 - t)^{a_k} = \alpha_k(t)$.

Step 4: the hull coordinate. Since $\mathbf{1}^\top \mathbf{y}^{\ell,k} = 1$, Eq. (16) gives

$$h^{\ell,k}(\varsigma) := \mathbf{1}^\top \mathbf{Z}^{\ell,k}(\varsigma) - 1 = \lambda_k \mathbf{1}^\top \mathbf{x}_0^{\ell,k} + (1 - \lambda_k) - 1 = \lambda_k(\varsigma) (\mathbf{1}^\top \mathbf{x}_0^{\ell,k} - 1), \quad (70)$$

in agreement with Proposition 3.4(ii). For $\mathbf{y} \in \Delta^{v_k-1}$, $0 \leq y_i \leq 1$ gives $\|\mathbf{y}\|^2 = \sum_i y_i^2 \leq \sum_i y_i = 1$. Hence, by the triangle inequality and $\lambda_k \in (0, 1]$,

$$\|\mathbf{Z}^{\ell,k}(\varsigma)\| \leq \lambda_k \|\mathbf{x}_0^{\ell,k}\| + (1 - \lambda_k) \|\mathbf{y}^{\ell,k}\| \leq \|\mathbf{x}_0^{\ell,k}\| + 1. \quad (71)$$

Step 5: $\varsigma_{\max} = \infty$. Fix a block (ℓ_0, k_0) with $h_0 := \mathbf{1}^\top \mathbf{x}_0^{\ell_0, k_0} - 1 \neq 0$ and suppose, for contradiction, $\varsigma_{\max} < \infty$. For $s < S$ we have $\varsigma(s) < \varsigma_{\max}$, so by Eq. (70)

$$|h^{\ell_0, k_0}(\varsigma(s))| = e^{-a_{k_0} \varsigma(s)} |h_0| \geq e^{-a_{k_0} \varsigma_{\max}} |h_0| =: \eta > 0.$$

The affine function $H(\mathbf{x}) := \mathbf{1}^\top \mathbf{x}^{\ell_0, k_0} - 1$ vanishes on $\mathcal{M}_1$ (as $\mathbf{1}^\top \mathbf{e}_i = 1$), and by Cauchy-Schwarz $|H(\mathbf{x}) - H(\mathbf{x}')| = |\mathbf{1}^\top (\mathbf{x} - \mathbf{x}')^{\ell_0, k_0}| \leq \sqrt{v_{k_0}} \|\mathbf{x} - \mathbf{x}'\|$. For $\mathbf{m} \in \mathcal{M}_1$ therefore $\eta \leq |H(\mathbf{X}_s) - H(\mathbf{m})| \leq \sqrt{v_{k_0}} \|\mathbf{X}_s - \mathbf{m}\|$, i.e. $\text{dist}(\mathbf{X}_s, \mathcal{M}_1) \geq \eta / \sqrt{v_{k_0}}$. By Eq. (71) and Minkowski's inequality in $\mathbb{R}^{LK}$,

$$\|\mathbf{X}_s\| = \left(\sum_{\ell, k} \|\mathbf{Z}^{\ell, k}\|^2 \right)^{1/2} \leq \left(\sum_{\ell, k} (\|\mathbf{x}_0^{\ell, k}\| + 1)^2 \right)^{1/2} \leq \left(\sum_{\ell, k} \|\mathbf{x}_0^{\ell, k}\|^2 \right)^{1/2} + \left(\sum_{\ell, k} 1 \right)^{1/2} = \|\mathbf{x}_0\| + \sqrt{LK}.$$

Hence $\mathbf{X}_s$ lies, for all $s \in [0, S)$, in $\mathcal{K} := \{\mathbf{x} : \|\mathbf{x}\| \leq \|\mathbf{x}_0\| + \sqrt{LK}, \text{dist}(\mathbf{x}, \mathcal{M}_1) \geq \eta / \sqrt{v_{k_0}}\}$, which is closed (both defining functions are continuous) and bounded, hence compact, and disjoint from $\mathcal{M}_1$, so $\mathcal{K} \subset \mathcal{O}$. A maximal forward solution of a locally Lipschitz ODE on an open set that stays in a compact subset of that set exists for all times, so $S = \infty$. By Eq. (66), $\varsigma(s) \geq s \rightarrow \infty$ as $s \rightarrow \infty$, contradicting $\varsigma(s) < \varsigma_{\max} < \infty$. Hence $\varsigma_{\max} = \infty$.

Under $\mu_0, \mathbf{x}_0^{\ell, k} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}_{v_k})$, so $\mathbf{1}^\top \mathbf{x}_0^{\ell, k}$ is a sum of v_k independent $\mathcal{N}(0, \sigma_k^2)$ variables, i.e. $\mathcal{N}(0, v_k \sigma_k^2)$ -distributed with $v_k \sigma_k^2 > 0$; a nondegenerate Gaussian has no atoms, so $\mathbb{P}(\mathbf{1}^\top \mathbf{x}_0^{\ell, k} = 1) = 0$ for every block. Also $\mu_0(\mathcal{M}_1) = 0$, as $\mathcal{M}_1$ is finite and μ_0 has a density.

Step 6: distance to the simplex. By Eq. (16), $\mathbf{Z}^{\ell, k} - \mathbf{y}^{\ell, k} = \lambda_k \mathbf{x}_0^{\ell, k} + (1 - \lambda_k) \mathbf{y}^{\ell, k} - \mathbf{y}^{\ell, k} = \lambda_k (\mathbf{x}_0^{\ell, k} - \mathbf{y}^{\ell, k})$. Since $\mathbf{y}^{\ell, k} \in \Delta^{v_k-1}$ and $\|\mathbf{y}^{\ell, k}\| \leq 1$ (Step 4),

$$\text{dist}(\mathbf{Z}^{\ell, k}, \Delta^{v_k-1}) \leq \|\mathbf{Z}^{\ell, k} - \mathbf{y}^{\ell, k}\| = \lambda_k \|\mathbf{x}_0^{\ell, k} - \mathbf{y}^{\ell, k}\| \leq \lambda_k (\|\mathbf{x}_0^{\ell, k}\| + 1),$$

for every $\varsigma \geq 0$ (as $\varsigma_{\max} = \infty$), and $\lambda_k(\varsigma) = e^{-a_k \varsigma} \rightarrow 0$ as $\varsigma \rightarrow \infty$. Since $\varsigma_{\max} = \infty$, $\mathbf{Z}(\varsigma) \in \mathcal{O}$ is defined for every finite ς ; moreover, for $\varsigma \leq \bar{\varsigma} < \infty$ the argument of Step 5 with $\varsigma_{\max}$ replaced by $\bar{\varsigma}$ gives $\text{dist}(\mathbf{Z}(\varsigma), \mathcal{M}_1) \geq e^{-a_{k_0} \bar{\varsigma}} |h_0| / \sqrt{v_{k_0}} > 0$. So the solution neither reaches nor accumulates at $\mathcal{M}_1$ at a finite flow time. $\square$

Proof of Corollary 4.8. *Idea.* Both parts are identities between closed-form expressions. Part (i) compares powers of $e^{-\varsigma}$. For part (ii) it suffices to observe that the flow-time ratio $\tilde{\rho}_k(\varsigma)$ equals the interpolant ratio $\rho_k(t)$ at $t = 1 - e^{-\varsigma}$, so its threshold crossing is that of Corollary 4.4 transported by the increasing map $t \mapsto -\ln(1 - t)$, which preserves the order of the levels.

(i) $\lambda_k = e^{-a_k \varsigma} = (e^{-a_{k'} \varsigma})^{a_k/a_{k'}} = \lambda_{k'}^{a_k/a_{k'}}$. If $\lambda_c = \varepsilon \in (0, 1)$, then $\lambda_f = \varepsilon^{a_f/a_c}$, and for $a_f < a_c$ the exponent $p := a_f/a_c$ lies in $(0, 1)$, so $\lambda_f/\varepsilon = \varepsilon^{p-1} > 1$ because $\varepsilon < 1$ and $p - 1 < 0$. By Eq. (16) the prior component of block (ℓ, k) is exactly $\lambda_k \mathbf{x}_0^{\ell, k}$, a fraction λ_k of the prior sample.

(ii) Multiplying numerator and denominator by $e^{a_k \varsigma}$,

$$\tilde{\rho}_k(\varsigma) = \frac{1 - e^{-a_k \varsigma}}{e^{-a_k \varsigma} \sigma_k} = \frac{e^{a_k \varsigma} - 1}{\sigma_k}.$$

By Corollary 4.4 with $a = a_k$, $\sigma = \sigma_k$, $\rho_k(t) = ((1 - t)^{-a_k} - 1)/\sigma_k$, and at $t = 1 - e^{-\varsigma}$, $(1 - t)^{-a_k} = e^{a_k \varsigma}$, so $\rho_k(1 - e^{-\varsigma}) = \tilde{\rho}_k(\varsigma)$. The map $\varsigma \mapsto \tilde{\rho}_k(\varsigma)$ is strictly increasing and continuous with $\tilde{\rho}_k(0) = 0$ and $\tilde{\rho}_k(\varsigma) \rightarrow \infty$; it equals $s_{v_k} = \sqrt{2 \log(v_k - 1)} \geq 0$ iff $e^{a_k \varsigma} = 1 + \sigma_k s_{v_k}$, i.e. iff $\varsigma = \ln(1 + \sigma_k s_{v_k})/a_k$, which is the first expression in Eq. (17). By Eq. (15), $1 - t_k^* = (1 + \sigma_k s_{v_k})^{-1/a_k}$, hence

$$-\ln(1 - t_k^*) = -\ln\left((1 + \sigma_k s_{v_k})^{-1/a_k}\right) = \frac{\ln(1 + \sigma_k s_{v_k})}{a_k} = \varsigma_k^*.$$

The map $t \mapsto -\ln(1 - t)$ is strictly increasing on $[0, 1)$, so $\varsigma_k^* < \varsigma_{k'}$ iff $t_k^* < t_{k'}^*$, and the order is that of Theorem 4.5. $\square$

Proof of Theorem 4.9. *Idea.* The order of two coordinates can change only if the trajectory crosses the hyperplane on which they are equal, so it suffices to show that this hyperplane is invariant. The tool is symmetry: exchanging the two symbols leaves the data, hence the field, unchanged (Step 1), so the

reflection of a solution is again a solution, and by uniqueness a solution that touches the fixed hyperplane of the reflection stays in it (Step 2). For product-uniform data this holds for every pair of symbols, which fixes the order of the coordinates at time zero; parts (ii) and (iii) follow because the targets are time averages of softmaxes, which preserve the order, and a vertex has a strictly largest coordinate.

Step 1: equivariance. Let $\mathbf{\Pi} \in \mathbb{R}^{v_k \times v_k}$ be the permutation matrix of the transposition ϖ of i and i' , so $\mathbf{\Pi} \mathbf{e}_m = \mathbf{e}_{\varpi(m)}$, and let $\mathbf{P} \in \mathbb{R}^{D \times D}$ act as $\mathbf{\Pi}$ on block (ℓ, k) and as the identity on all other blocks. Let $\theta : \mathcal{J} \rightarrow \mathcal{J}$ replace the entry $j_{\ell, k}$ by $\varpi(j_{\ell, k})$. Then $\mathbf{P}$ is orthogonal and symmetric, θ is an involution of $\mathcal{J}$, $\mathbf{P} \mathbf{e}_j = \mathbf{e}_{\theta(j)}$ (block (ℓ, k) : $\mathbf{\Pi} \mathbf{e}_{j_{\ell, k}} = \mathbf{e}_{\varpi(j_{\ell, k})}$; other blocks unchanged), hence $\mathbf{P} \mathcal{M}_1 = \mathcal{M}_1$ and $\mathbf{P} \mathcal{O} = \mathcal{O}$. With $\mathbf{D}(t) := \text{diag}(\frac{\alpha_{k'}}{1-t} \mathbf{I}_{v_{k'}})_{\ell', k'}$, $\mathbf{P} \mathbf{D}(t) = \mathbf{D}(t) \mathbf{P}$, because $\mathbf{D}(t)$ is a scalar multiple of the identity on each block and $\mathbf{P}$ maps each block to itself.

The likelihood is invariant. For a block $(\ell', k') \neq (\ell, k)$, $(\mathbf{P} \mathbf{x})^{\ell', k'} = \mathbf{x}^{\ell', k'}$ and $\theta(j)_{\ell', k'} = j_{\ell', k'}$; for block (ℓ, k) , by orthogonality of $\mathbf{\Pi}$,

$$\|(\mathbf{P} \mathbf{x})^{\ell, k} - (1 - \alpha_k) \mathbf{e}_{\theta(j)_{\ell, k}}\| = \|\mathbf{\Pi}(\mathbf{x}^{\ell, k} - (1 - \alpha_k) \mathbf{e}_{j_{\ell, k}})\| = \|\mathbf{x}^{\ell, k} - (1 - \alpha_k) \mathbf{e}_{j_{\ell, k}}\|.$$

Since Eq. (4) depends on $\mathbf{x}$ and $\mathbf{j}$ only through these norms, $p(\mathbf{P} \mathbf{x} | t, \theta(\mathbf{j})) = p(\mathbf{x} | t, \mathbf{j})$. The invariance of μ_1 means $\sum_j \pi_j \delta_{\mathbf{P} \mathbf{e}_j} = \sum_j \pi_j \delta_{\mathbf{e}_j}$; as $\mathbf{P} \mathbf{e}_j = \mathbf{e}_{\theta(j)}$ and the vertices are distinct, $\pi_{\theta(j)} = \pi_j$. Reindexing the sum by the bijection $\mathbf{j} \mapsto \theta(\mathbf{j})$,

$$q(\mathbf{P} \mathbf{x}) = \int_0^1 \sum_j \pi_{\theta(j)} p(\mathbf{P} \mathbf{x} | t, \theta(\mathbf{j})) dt = \int_0^1 \sum_j \pi_j p(\mathbf{x} | t, \mathbf{j}) dt = q(\mathbf{x}),$$

and therefore $w_{\theta(j)}(t, \mathbf{P} \mathbf{x}) = \pi_{\theta(j)} p(\mathbf{P} \mathbf{x} | t, \theta(\mathbf{j})) / q(\mathbf{P} \mathbf{x}) = w_j(t, \mathbf{x})$. In matrix form Eq. (5) reads $\mathbf{b}(\mathbf{x}) = \int_0^1 \sum_j w_j(t, \mathbf{x}) \mathbf{D}(t) (\mathbf{e}_j - \mathbf{x}) dt$ (block (ℓ', k') of $\mathbf{D}(t) (\mathbf{e}_j - \mathbf{x})$ is $\frac{\alpha_{k'}}{1-t} (\mathbf{e}_{j_{\ell', k'}} - \mathbf{x}^{\ell', k'})$). For $\mathbf{x} \in \mathcal{O}$, reindexing $\mathbf{j} \mapsto \theta(\mathbf{j})$, then using $w_{\theta(j)}(t, \mathbf{P} \mathbf{x}) = w_j(t, \mathbf{x})$, $\mathbf{e}_{\theta(j)} = \mathbf{P} \mathbf{e}_j$ and $\mathbf{D} \mathbf{P} = \mathbf{P} \mathbf{D}$, and finally pulling the constant matrix $\mathbf{P}$ out of the absolutely convergent integral,

$$\mathbf{b}(\mathbf{P} \mathbf{x}) = \int_0^1 \sum_j w_{\theta(j)}(t, \mathbf{P} \mathbf{x}) \mathbf{D}(t) (\mathbf{e}_{\theta(j)} - \mathbf{P} \mathbf{x}) dt = \int_0^1 \sum_j w_j(t, \mathbf{x}) \mathbf{P} \mathbf{D}(t) (\mathbf{e}_j - \mathbf{x}) dt = \mathbf{P} \mathbf{b}(\mathbf{x}).$$

Step 2: invariance of the hyperplane. Let $H := \{\mathbf{x} : x_i^{\ell, k} = x_{i'}^{\ell, k}\}$; then $H = \{\mathbf{x} : \mathbf{P} \mathbf{x} = \mathbf{x}\}$. Let $\mathbf{X} : (s_-, s_+) \rightarrow \mathcal{O}$ be the maximal (two-sided) solution through a point $\mathbf{X}_{s_1} \in H \cap \mathcal{O}$. Then $\tilde{\mathbf{X}}_s := \mathbf{P} \mathbf{X}_s \in \mathbf{P} \mathcal{O} = \mathcal{O}$, and by Step 1 $\tilde{\mathbf{X}}_s = \mathbf{P} \mathbf{b}(\mathbf{X}_s) = \mathbf{b}(\mathbf{P} \mathbf{X}_s) = \mathbf{b}(\tilde{\mathbf{X}}_s)$, with $\tilde{\mathbf{X}}_{s_1} = \mathbf{P} \mathbf{X}_{s_1} = \mathbf{X}_{s_1}$. By uniqueness for the locally Lipschitz field $\mathbf{b}$ on $\mathcal{O}$, $\tilde{\mathbf{X}} = \mathbf{X}$ on (s_-, s_+) , i.e. $\mathbf{X}_s \in H$ for all $s \in (s_-, s_+)$. Now let $\mathbf{X} : [0, S) \rightarrow \mathcal{O}$ be a solution with $\mathbf{x}_0 \notin H$. If $\mathbf{X}_{s_1} \in H$ for some s_1 , the two-sided maximal solution through $\mathbf{X}_{s_1}$ extends $\mathbf{X}$ (uniqueness), so it passes through $\mathbf{x}_0$ at time 0, and $\mathbf{x}_0 \in H$ by what was just shown, a contradiction. Hence the continuous function $s \mapsto X_{s, i}^{\ell, k} - X_{s, i'}^{\ell, k}$ never vanishes on $[0, S)$, and by the intermediate value theorem its sign is constant. If $\mathbf{x}_0 \in H$, it vanishes identically.

Step 3: product-uniform data. Then $\pi_j = \prod_{\ell, k} v_k^{-1}$ does not depend on $\mathbf{j}$, so μ_1 is invariant under every exchange of two symbols at every block, and Step 2 applies to all blocks and all pairs $i \neq i'$. Under μ_0 , $x_{0, i}^{\ell, k} - x_{0, i'}^{\ell, k} \sim \mathcal{N}(0, 2\sigma_k^2)$ has no atom at 0, and there are finitely many triples (ℓ, k, i, i') , so almost surely the coordinates of every block of $\mathbf{x}_0$ are pairwise distinct; also $\mathbf{x}_0 \notin \mathcal{M}_1$ almost surely. For such $\mathbf{x}_0$, for every pair $i \neq i'$ the sign of $X_{s, i}^{\ell, k} - X_{s, i'}^{\ell, k}$ is that of $x_{0, i}^{\ell, k} - x_{0, i'}^{\ell, k} \neq 0$ for all $s \in [0, S)$, so the strict order of the coordinates of every block is preserved, which is (i).

For (ii), let $\mathbf{x} \in \mathcal{O}$ and $t \in (0, 1)$, and let $Z_t(\mathbf{x}) := \sum_j \pi_j p(\mathbf{x} | t, \mathbf{j}) > 0$ be the normaliser of Eq. (54). With the time posterior $\nu(t | \mathbf{x}) := Z_t(\mathbf{x}) / q(\mathbf{x})$, which is positive and finite (Lemma B.3(ii)),

$$w_j(t, \mathbf{x}) = \frac{\pi_j p(\mathbf{x} | t, \mathbf{j})}{q(\mathbf{x})} = \frac{Z_t(\mathbf{x})}{q(\mathbf{x})} \cdot \frac{\pi_j p(\mathbf{x} | t, \mathbf{j})}{Z_t(\mathbf{x})} = \nu(t | \mathbf{x}) \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j | \mathbf{I}_t = \mathbf{x}, t).$$

By Proposition 4.1, $\mathbb{P}(\mathbf{x}_1 = \mathbf{e}_j | \mathbf{I}_t = \mathbf{x}, t) = \prod_b q^b(j_b)$ with $q^{\ell, k}(i) = \text{softmax}(\beta_k(t) \mathbf{x}^{\ell, k})_i$, and by Eq. (57) applied to the blocks other than (ℓ, k) , each of whose factors sums to one, $\sum_{\mathbf{j} : j_{\ell, k} = i} \prod_b q^b(j_b) = q^{\ell, k}(i)$. Summing Eq. (6) over the configurations with $j_{\ell, k} = i$ (a finite sum of nonnegative integrable functions, so

sum and integral commute),

$$\bar{p}_i^{\ell,k}(\mathbf{x}) = \frac{1}{M(\mathbf{x})} \int_0^1 \frac{1}{1-t} \sum_{j: j_{\ell,k}=i} w_j(t, \mathbf{x}) dt = \frac{1}{M(\mathbf{x})} \int_0^1 \frac{\nu(t | \mathbf{x})}{1-t} \frac{\exp(\beta_k(t)x_i^{\ell,k})}{\sum_{i''} \exp(\beta_k(t)x_{i''}^{\ell,k})} dt.$$

For $t \in (0, 1)$, $\alpha_k(t) \in (0, 1)$ and so $\beta_k(t) = (1 - \alpha_k)/(\alpha_k^2 \sigma_k^2) > 0$. If $x_i^{\ell,k} > x_{i'}^{\ell,k}$, then $e^{\beta_k x_i^{\ell,k}} > e^{\beta_k x_{i'}^{\ell,k}}$, so

$$\bar{p}_i^{\ell,k}(\mathbf{x}) - \bar{p}_{i'}^{\ell,k}(\mathbf{x}) = \frac{1}{M(\mathbf{x})} \int_0^1 \frac{\nu(t | \mathbf{x})}{1-t} \frac{e^{\beta_k(t)x_i^{\ell,k}} - e^{\beta_k(t)x_{i'}^{\ell,k}}}{\sum_{i''} e^{\beta_k(t)x_{i''}^{\ell,k}}} dt > 0,$$

because the integrand is strictly positive on $(0, 1)$ and the integral is finite (at most $\int_0^1 \nu/(1-t) dt = M(\mathbf{x})$). By (i), $X_{s,i^*}^{\ell,k} > X_{s,i'}^{\ell,k}$ for all $i' \neq i^* := i_{\ell,k}^*$ and all s , so $\bar{\mathbf{p}}^{\ell,k}(\mathbf{X}_s)$ has its strictly largest coordinate at i^* for all s . Hence $\mathbf{f}(\varsigma') := \bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma'))$ satisfies $f_{i^*} - f_{i'} > 0$ on $[0, \varsigma_{\max})$. For $\varsigma = 0$, $\mathbf{y}^{\ell,k}(0) = \mathbf{f}(0)$ (Step 3 of the proof of [Theorem 4.7](#)); for $\varsigma > 0$,

$$y_{i^*}^{\ell,k}(\varsigma) - y_{i'}^{\ell,k}(\varsigma) = \frac{1}{1 - \lambda_k(\varsigma)} \int_0^\varsigma \kappa_\varsigma(\varsigma') (f_{i^*}(\varsigma') - f_{i'}(\varsigma')) d\varsigma' > 0,$$

as the integrand is continuous and strictly positive on $[0, \varsigma]$.

For (iii), let $\mathbf{X}_s \rightarrow \mathbf{e}_{j^*} \in \mathcal{M}_1$ as $s \uparrow S$ and suppose $m := j_{\ell,k}^* \neq i_{\ell,k}^*$ for some block. By (i), $X_{s,i^*}^{\ell,k} - X_{s,m}^{\ell,k} > 0$ for all s , so the limit satisfies $(\mathbf{e}_m)_{i^*} - (\mathbf{e}_m)_m \geq 0$, i.e. $0 - 1 \geq 0$, a contradiction.

Finally, fix a block. For a permutation matrix $\mathbf{\Pi}$ of $[v_k]$, $\mathbf{\Pi} \mathbf{x}_0^{\ell,k}$ has the same law $\mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}_{v_k})$ as $\mathbf{x}_0^{\ell,k}$, and the almost surely unique argmax satisfies $\arg \max_m (\mathbf{\Pi} \mathbf{x})_m = \varpi(\arg \max_m x_m)$ for the permutation ϖ of $\mathbf{\Pi}$. Hence $\mathbb{P}(\arg \max_m x_{0,m}^{\ell,k} = i)$ does not depend on i , and as these probabilities sum to one, each equals $1/v_k$. Since $\mathbf{\Sigma}$ is block diagonal, the Gaussian blocks of $\mathbf{x}_0$ are independent, so the blockwise argmax map $\mathbf{x} \mapsto (e_{\arg \max_m x_m^{\ell,k}})_{\ell,k}$ pushes μ_0 to $\bigotimes_{\ell,k} \text{Unif}([v_k]) = \mu_1$. For $\mathbf{x}_0 \in \mathcal{A}$ in the full-measure set of Step 3, $\mathbf{T}(\mathbf{x}_0) = \lim_{s \uparrow \tau(\mathbf{x}_0)} \mathbf{X}_s \in \mathcal{M}_1^+ \subset \mathcal{M}_1$ ([Theorem 3.8](#); the solution in $\Omega = \mathcal{O}$ is the one considered here), so (iii) identifies $\mathbf{T}(\mathbf{x}_0)$ with the blockwise argmax. $\square$

Proof of Theorem 4.10. *Idea.* Both bounds come from the variation-of-constants formula [Eq. \(68\)](#). For the lower bound we need a quantity that the flow cannot reduce faster than the schedule; the hull coordinate $\mathbf{1}^\top \mathbf{Z}^{\ell,k} - 1$ is one, since it decays exactly like λ_k by [Eq. \(70\)](#), and by Cauchy–Schwarz it bounds the distance to any vertex from below. For the upper bound we subtract the vertex from [Eq. \(68\)](#) and bound the remaining integral by $\lambda K_{\ell,k}$. Solving $\lambda_k(\varsigma) c = \varepsilon$ for ς turns both bounds into bounds on the arrival time, and the order carries over to the original time because $s(\cdot)$ is increasing.

Write $\mathbf{e}^* = \mathbf{e}_{\ell,k}^*$, $\lambda = \lambda_k(\varsigma)$, $\mathbf{f}(\varsigma') = \bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma'))$ and κ_ς as in [Eq. \(69\)](#). The almost-sure conditions of [Theorem 4.7](#) give $\varsigma_{\max} = \infty$, so [Eqs. \(68\)](#) and [\(70\)](#) hold for all $\varsigma \geq 0$; those of [Theorem 4.9](#) make $i_{\ell,k}^*$, hence $\mathbf{e}^*$, well defined.

Lower bound. By Cauchy–Schwarz, $|\mathbf{1}^\top \mathbf{u}| \leq \|\mathbf{1}\| \|\mathbf{u}\| = \sqrt{v_k} \|\mathbf{u}\|$ for $\mathbf{u} \in \mathbb{R}^{v_k}$. Take $\mathbf{u} = \mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}^*$. As $\mathbf{1}^\top \mathbf{e}^* = 1$, [Eq. \(70\)](#) gives $\mathbf{1}^\top \mathbf{u} = \mathbf{1}^\top \mathbf{Z}^{\ell,k}(\varsigma) - 1 = \lambda(\mathbf{1}^\top \mathbf{x}_0^{\ell,k} - 1)$, so $\lambda|\mathbf{1}^\top \mathbf{x}_0^{\ell,k} - 1| \leq \sqrt{v_k} \|\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}^*\|$, the left inequality of [Eq. \(19\)](#).

Upper bound. By [Eq. \(69\)](#), $\mathbf{e}^* = \lambda \mathbf{e}^* + (1 - \lambda) \mathbf{e}^* = \lambda \mathbf{e}^* + \int_0^\varsigma \kappa_\varsigma(\varsigma') \mathbf{e}^* d\varsigma'$. Subtracting this from [Eq. \(68\)](#),

$$\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}^* = \lambda(\mathbf{x}_0^{\ell,k} - \mathbf{e}^*) + \int_0^\varsigma a_k e^{-a_k(\varsigma - \varsigma')} (\mathbf{f}(\varsigma') - \mathbf{e}^*) d\varsigma'.$$

For a continuous $\mathbf{g} : [0, \varsigma] \rightarrow \mathbb{R}^{v_k}$ and $\mathbf{v} := \int \mathbf{g} \neq \mathbf{0}$, $\|\mathbf{v}\|^2 = \int \langle \mathbf{v}, \mathbf{g} \rangle \leq \|\mathbf{v}\| \int \|\mathbf{g}\|$, so $\|\int \mathbf{g}\| \leq \int \|\mathbf{g}\|$. Hence, as the integrand of [Eq. \(18\)](#) is nonnegative,

$$\left\| \int_0^\varsigma a_k e^{-a_k(\varsigma - \varsigma')} (\mathbf{f} - \mathbf{e}^*) d\varsigma' \right\| \leq e^{-a_k \varsigma} \int_0^\varsigma a_k e^{a_k \varsigma'} \|\mathbf{f}(\varsigma') - \mathbf{e}^*\| d\varsigma' \leq e^{-a_k \varsigma} \int_0^\infty a_k e^{a_k \varsigma'} \|\mathbf{f}(\varsigma') - \mathbf{e}^*\| d\varsigma' = \lambda K_{\ell,k}.$$

The triangle inequality gives $\|\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}^*\| \leq \lambda \|\mathbf{x}_0^{\ell,k} - \mathbf{e}^*\| + \lambda K_{\ell,k}$, the right inequality of [Eq. \(19\)](#) (trivial if $K_{\ell,k} = \infty$).

Arrival times. Write $B := |\mathbf{1}^\top \mathbf{x}_0^{\ell,k} - 1|/\sqrt{v_k}$ and $A := \|\mathbf{x}_0^{\ell,k} - \mathbf{e}^*\| + K_{\ell,k}$. Let $\varsigma \geq 0$ satisfy $\|\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}^*\| \leq \varepsilon$. By the lower bound, $e^{-a_k \varsigma} B \leq \varepsilon$. If $B = 0$ the left inequality of Eq. (20) reads $-\infty \leq \varsigma_{\ell,k}^\varepsilon$; otherwise taking logarithms gives $-a_k \varsigma + \ln B \leq \ln \varepsilon$, i.e. $\varsigma \geq \frac{1}{a_k} \ln \frac{B}{\varepsilon}$. Taking the infimum over all such ς gives the left inequality (if there is no such ς , $\varsigma_{\ell,k}^\varepsilon = \infty$).

For the right inequality let $K_{\ell,k} < \infty$ (if $K_{\ell,k} = \infty$ the right side is ∞). If $A \geq \varepsilon$, then $\varsigma_0 := \frac{1}{a_k} \ln \frac{A}{\varepsilon} \geq 0$, $\lambda_k(\varsigma_0) = \varepsilon/A$, and the upper bound gives $\|\mathbf{Z}^{\ell,k}(\varsigma_0) - \mathbf{e}^*\| \leq \lambda_k(\varsigma_0)A = \varepsilon$, so $\varsigma_{\ell,k}^\varepsilon \leq \varsigma_0$. If $A < \varepsilon$, then $\|\mathbf{Z}^{\ell,k}(0) - \mathbf{e}^*\| = \|\mathbf{x}_0^{\ell,k} - \mathbf{e}^*\| \leq A < \varepsilon$, so $\varsigma_{\ell,k}^\varepsilon = 0$. In all cases, which is the right inequality of Eq. (20),

$$\varsigma_{\ell,k}^\varepsilon \leq \max \left\{ 0, \frac{1}{a_k} \ln \frac{\|\mathbf{x}_0^{\ell,k} - \mathbf{e}_{\ell,k}^*\| + K_{\ell,k}}{\varepsilon} \right\}. \quad (72)$$

Order. Fix two levels c, f of position ℓ , let $U_c := \frac{1}{a_c} \ln \frac{\|\mathbf{x}_0^{\ell,c} - \mathbf{e}_{\ell,c}^*\| + K_{\ell,c}}{\varepsilon}$ and $L_f := \frac{1}{a_f} \ln \frac{|\mathbf{1}^\top \mathbf{x}_0^{\ell,f} - 1|}{\varepsilon \sqrt{v_f}}$, and assume, as in the statement, that the upper bound for c is smaller than the lower bound for f : $\max\{0, U_c\} < L_f$. In particular $L_f > 0$. By Eq. (72) and the left inequality of Eq. (20),

$$\varsigma_{\ell,c}^\varepsilon \leq \max\{0, U_c\} \quad \text{and} \quad L_f \leq \varsigma_{\ell,f}^\varepsilon.$$

Hence $\varsigma_{\ell,c}^\varepsilon \leq \max\{0, U_c\} < L_f \leq \varsigma_{\ell,f}^\varepsilon$. Both arrival times refer to the same trajectory. Since $\varsigma \mapsto s(\varsigma)$ is a continuous strictly increasing bijection of $[0, \infty)$ onto $[0, S)$ (Theorem 4.7), $\{s \in [0, S) : \|\mathbf{X}_s^{\ell,k} - \mathbf{e}_{\ell,k}^*\| \leq \varepsilon\} = s(\{\varsigma \geq 0 : \|\mathbf{Z}^{\ell,k}(\varsigma) - \mathbf{e}_{\ell,k}^*\| \leq \varepsilon\})$. The latter set is closed in $[0, \infty)$ by continuity of $\mathbf{Z}$, so when nonempty its infimum $\varsigma_{\ell,k}^\varepsilon$ is attained, and the first time in s at which block (ℓ, k) is within ε of its vertex is $s(\varsigma_{\ell,k}^\varepsilon)$ (and there is no such time if $\varsigma_{\ell,k}^\varepsilon = \infty$). As $s(\cdot)$ is strictly increasing, $s(\varsigma_{\ell,c}^\varepsilon) < s(\varsigma_{\ell,f}^\varepsilon)$ (with $s(\infty) := \infty$). $\square$

Proof of Corollary 4.11. *Idea.* By the order part of Theorem 4.10 it suffices that $\max\{0, U_c\} < L_f$, and Eq. (21) is chosen so that this holds whenever $|\mathbf{1}^\top \mathbf{x}_0^{\ell,f} - 1| \geq r$, $\|\mathbf{x}_0^{\ell,c}\| \leq R$ and $K_{\ell,c} \leq \bar{K}$. The first two events concern Gaussian quantities and are bounded directly, the first by the maximal density of a one-dimensional Gaussian and the second by Markov's inequality; the probability of the third is kept as it stands.

Consider the events $E_1 = \{|\mathbf{1}^\top \mathbf{x}_0^{\ell,f} - 1| \geq r\}$, $E_2 = \{\|\mathbf{x}_0^{\ell,c}\| \leq R\}$ and $E_3 = \{K_{\ell,c} \leq \bar{K}\}$, and let E_0 be the full-measure event of the almost-sure conditions of Theorem 4.10. The variable $G := \mathbf{1}^\top \mathbf{x}_0^{\ell,f} - 1$ is $\mathcal{N}(-1, v_f \sigma_f^2)$ -distributed (Step 5 of the proof of Theorem 4.7), and its density is at most $(2\pi v_f \sigma_f^2)^{-1/2}$, so

$$\mathbb{P}(E_1^c) = \mathbb{P}(-r < G < r) \leq 2r (2\pi v_f \sigma_f^2)^{-1/2} = \frac{2r}{\sqrt{2\pi v_f} \sigma_f}.$$

By Markov's inequality applied to $\|\mathbf{x}_0^{\ell,c}\|^2$ and $\mathbb{E}\|\mathbf{x}_0^{\ell,c}\|^2 = \sum_{i=1}^{v_c} \mathbb{E}(x_{0,i}^{\ell,c})^2 = v_c \sigma_c^2$, $\mathbb{P}(E_2^c) \leq \mathbb{P}(\|\mathbf{x}_0^{\ell,c}\|^2 \geq R^2) \leq v_c \sigma_c^2 / R^2$. On $E_0 \cap E_1 \cap E_2 \cap E_3$, using $\|\mathbf{x}_0^{\ell,c} - \mathbf{e}_{\ell,c}^*\| \leq \|\mathbf{x}_0^{\ell,c}\| + 1 \leq R + 1$ and the monotonicity of $\ln$,

$$U_c = \frac{1}{a_c} \ln \frac{\|\mathbf{x}_0^{\ell,c} - \mathbf{e}_{\ell,c}^*\| + K_{\ell,c}}{\varepsilon} \leq \frac{1}{a_c} \ln \frac{R + 1 + \bar{K}}{\varepsilon} < \frac{1}{a_f} \ln \frac{r}{\varepsilon \sqrt{v_f}} \leq \frac{1}{a_f} \ln \frac{|\mathbf{1}^\top \mathbf{x}_0^{\ell,f} - 1|}{\varepsilon \sqrt{v_f}} = L_f,$$

where the strict inequality is Eq. (21) multiplied by the number $\ln(r/(\varepsilon \sqrt{v_f}))/a_c$, which is positive because $r > \varepsilon \sqrt{v_f}$. For the same reason $L_f \geq \frac{1}{a_f} \ln \frac{r}{\varepsilon \sqrt{v_f}} > 0$. By the order part of Theorem 4.10, applicable because $U_c < L_f$ and $L_f > 0$ give $\max\{0, U_c\} < L_f$, $\varsigma_{\ell,c}^\varepsilon < \varsigma_{\ell,f}^\varepsilon$ on $E_0 \cap E_1 \cap E_2 \cap E_3$. Since $\mathbb{P}(E_0^c) = 0$, the union bound gives

$$\mathbb{P}(\varsigma_{\ell,c}^\varepsilon < \varsigma_{\ell,f}^\varepsilon) \geq 1 - \mathbb{P}(E_1^c) - \mathbb{P}(E_2^c) - \mathbb{P}(E_3^c) \geq 1 - \frac{2r}{\sqrt{2\pi v_f} \sigma_f} - \frac{v_c \sigma_c^2}{R^2} - \mathbb{P}(K_{\ell,c} > \bar{K}). \quad \square$$

Proof of Proposition 4.12. *Idea.* The integrand of Eq. (18) is $a_k e^{a_k \varsigma} \|\bar{\mathbf{p}}^{\ell,k} - \mathbf{e}_{\ell,k}^*\|$, so $K_{\ell,k} < \infty$ once the target approaches its vertex faster than $e^{-a_k \varsigma}$. Near the limit vertex Lemma B.4(d) bounds this distance by $C_0 V^N$, with V the schedule-weighted distance Eq. (8) and $N > a_{\max}$, so it suffices to show that V decays like $e^{-\varsigma}$ in flow time. This is the differential inequality behind Theorem 3.7, rewritten in flow time, where the speed M drops out (Step 1); Steps 2 and 3 insert the decay into the bound for the targets and into $K_{\ell,k}$.

Let $\mathbf{x}_0$ satisfy the almost-sure conclusions of [Theorems 3.8, 4.7 and 4.9](#) (a countable intersection of full-measure events). Since $\Omega = \mathcal{O}$ for product-uniform data, the maximal solution of [Theorem 4.7](#) is the one of [Theorem 3.8](#), so $S = \tau(\mathbf{x}_0) < \infty$ and $\mathbf{X}_s \rightarrow \mathbf{e}_{j^*} := \mathbf{T}(\mathbf{x}_0) \in \mathcal{M}_1^+$ as $s \uparrow \tau(\mathbf{x}_0)$. By [Theorem 4.9\(iii\)](#), $j_{\ell,k}^* = i_{\ell,k}^*$ for every block, so $\mathbf{e}_{\ell,k}^*$ in [Eq. \(18\)](#) is block (ℓ, k) of the limit. Let ε, C_0, γ be as in the proof of [Theorem 3.7](#), so $\varepsilon \leq \varepsilon_0 < 1$ and $C_0 \varepsilon^\gamma \leq \frac{1}{2}$. As V_{j^*} is continuous with $V_{j^*}(\mathbf{e}_{j^*}) = 0$, there is $s_1 < \tau(\mathbf{x}_0)$ with $V_{j^*}(\mathbf{X}_{s_1}) < \varepsilon$; and $V_{j^*}(\mathbf{X}_{s_1}) > 0$ because $\mathbf{X}_{s_1} \in \mathcal{O}$ is not the vertex. So $\mathbf{X}_{s_1} \in W = \{0 < V_{j^*} < \varepsilon\}$. By uniqueness, $s \mapsto \mathbf{X}_{s_1+s}$ is the solution started at $\mathbf{X}_{s_1}$, so by [Theorem 3.7\(i\)](#) $\mathbf{X}_s \in W$ for all $s \in [s_1, \tau(\mathbf{x}_0))$. Put $\varsigma_1 := \varsigma(s_1)$, $\tilde{V}(s) := V_{j^*}(\mathbf{X}_s)$ and $V(\varsigma) := V_{j^*}(\mathbf{Z}(\varsigma)) = \tilde{V}(s(\varsigma))$. By [Theorem 4.7](#), $s(\cdot)$ maps $[\varsigma_1, \infty)$ onto $[s_1, \tau(\mathbf{x}_0))$, so $0 < V(\varsigma) < \varepsilon$ for $\varsigma \geq \varsigma_1$.

Step 1: decay in flow time. Let $s \in [s_1, \tau(\mathbf{x}_0))$. In Step 1 of the proof of [Theorem 3.7](#), an active block $(f_{\ell,k}(s) = \tilde{V}(s) > 0)$, so $f_{\ell,k}$ is C^1 near s) satisfies, with $\|\delta^{\ell,k}\|^{1/a_k-1} = \tilde{V}^{1-a_k}$,

$$f'_{\ell,k}(s) \leq M(\mathbf{X}_s)(-\tilde{V} + C_0 \tilde{V}^{1-a_k} \tilde{V}^N) \leq M(\mathbf{X}_s)(-\tilde{V} + C_0 \tilde{V}^{1+\gamma}),$$

because $1 - a_k + N = 1 + (N - a_k) \geq 1 + (N - a_{\max}) = 1 + \gamma$ and $0 < \tilde{V} < 1$. The inactive $f_{\ell,k}$ are continuous, which is all the proof of [Lemma B.5](#) uses about them, so [Lemma B.5](#) gives $D^+ \tilde{V}(s) \leq M(\mathbf{X}_s)(-\tilde{V}(s) + C_0 \tilde{V}(s)^{1+\gamma}) < \infty$. By [Lemma C.9\(a\)](#) with $\sigma = s(\cdot)$ and $s'(\varsigma) = 1/M(\mathbf{Z}(\varsigma))$ ([Eq. \(67\)](#)), and since $C_0 V^\gamma \leq C_0 \varepsilon^\gamma \leq \frac{1}{2}$,

$$D_\varsigma^+ V \leq -V(1 - C_0 V^\gamma) \leq -\frac{1}{2}V \quad (\varsigma \geq \varsigma_1). \quad (73)$$

By [Lemma C.9\(b\)](#) with $\phi = \ln$ (which is C^1 with $\phi' > 0$ on $(0, \infty) \ni V(\varsigma)$), $D^+ \ln V = D^+ V/V \leq -\frac{1}{2}$, and [Lemma B.5](#) with $\kappa = \frac{1}{2}$ on every interval $[\varsigma_1, \varsigma_2]$ gives $\ln V(\varsigma) \leq \ln V(\varsigma_1) - \frac{1}{2}(\varsigma - \varsigma_1)$, i.e. $V(\varsigma) \leq V(\varsigma_1)e^{-(\varsigma - \varsigma_1)/2}$. Hence

$$B := \int_{\varsigma_1}^\infty C_0 V^\gamma d\varsigma \leq C_0 V(\varsigma_1)^\gamma \int_{\varsigma_1}^\infty e^{-\gamma(\varsigma - \varsigma_1)/2} d\varsigma = \frac{2C_0 V(\varsigma_1)^\gamma}{\gamma} < \infty.$$

Now let $F(\varsigma) := \ln V(\varsigma) + (\varsigma - \varsigma_1) - \int_{\varsigma_1}^\varsigma C_0 V^\gamma$. The last two terms are differentiable with derivative $1 - C_0 V(\varsigma)^\gamma$ (fundamental theorem of calculus; V is continuous). By the first inequality in [Eq. \(73\)](#) and [Lemma C.9\(b\)](#), $D^+ \ln V = D^+ V/V \leq -1 + C_0 V^\gamma$, so by [Lemma C.9\(c\)](#)

$$D^+ F \leq (-1 + C_0 V^\gamma) + (1 - C_0 V^\gamma) = 0.$$

[Lemma B.5](#) with $\kappa = 0$ gives $F(\varsigma) \leq F(\varsigma_1) = \ln V(\varsigma_1)$, i.e. $\ln V(\varsigma) \leq \ln V(\varsigma_1) - (\varsigma - \varsigma_1) + \int_{\varsigma_1}^\varsigma C_0 V^\gamma \leq \ln V(\varsigma_1) - (\varsigma - \varsigma_1) + B$. Exponentiating,

$$V(\varsigma) \leq V(\varsigma_1) e^B e^{-(\varsigma - \varsigma_1)} =: C_1 e^{-\varsigma} \quad (\varsigma \geq \varsigma_1), \quad C_1 := V(\varsigma_1) e^{B + \varsigma_1} < \infty.$$

Step 2: the targets. For $\varsigma \geq \varsigma_1$, $0 < V(\varsigma) < \varepsilon \leq \varepsilon_0$, so [Lemma B.4\(d\)](#) and Step 1 give $\|\bar{\mathbf{p}}^{\ell,k}(\mathbf{Z}(\varsigma)) - \mathbf{e}_{\ell,k}^*\| \leq C_0 V(\varsigma)^N \leq C_0 C_1^N e^{-N\varsigma}$. With $C := \max\{C_1, C_0 C_1^N\}$ this proves the two displayed bounds of the statement. Moreover $N = L \sum_k a_k v_k \geq v_k a_k \geq 2a_k$ for every k (as $v_k \geq 2$), so $N \geq 2a_{\max}$.

Step 3: K . The integrand of [Eq. \(18\)](#) is continuous and nonnegative; split the integral at ς_1 . For $\mathbf{u}, \mathbf{v} \in \Delta^{v_k-1}$, $|u_i - v_i| \leq 1$ and $\sum_i |u_i - v_i| \leq 2$, so $\|\mathbf{u} - \mathbf{v}\|^2 \leq \sum_i |u_i - v_i| \leq 2$; as $\bar{\mathbf{p}}^{\ell,k}$ and $\mathbf{e}_{\ell,k}^*$ lie in Δ^{v_k-1} , the integrand is at most $\sqrt{2}a_k e^{a_k \varsigma} \leq \sqrt{2}a_k e^{a_k \varsigma_1}$ on $[0, \varsigma_1]$. On $[\varsigma_1, \infty)$, by Step 2, it is at most $a_k e^{a_k \varsigma} C_0 C_1^N e^{-N\varsigma} = a_k C_0 C_1^N e^{-(N-a_k)\varsigma} \leq a_k C_0 C_1^N e^{-\gamma\varsigma}$, because $N - a_k \geq N - a_{\max} = \gamma > 0$ by [Eq. \(46\)](#) and $\varsigma \geq 0$. Hence

$$K_{\ell,k} \leq \sqrt{2}a_k \varsigma_1 e^{a_k \varsigma_1} + a_k C_0 C_1^N \int_{\varsigma_1}^\infty e^{-\gamma\varsigma} d\varsigma = \sqrt{2}a_k \varsigma_1 e^{a_k \varsigma_1} + \frac{a_k C_0 C_1^N e^{-\gamma\varsigma_1}}{\gamma} < \infty. \quad \square$$

Proof of Corollary 4.13. *Idea.* By the order part of [Theorem 4.10](#) it suffices that $U_c < L_f$ and $L_f > 0$ for all small ε . Up to constants that depend on $\mathbf{x}_0$, U_c is $\frac{1}{a_c} \ln \frac{1}{\varepsilon}$ and L_f is $\frac{1}{a_f} \ln \frac{1}{\varepsilon}$, two linear functions of $\ln \frac{1}{\varepsilon}$ whose slopes satisfy $1/a_f > 1/a_c$, so the comparison holds for small ε as soon as the constants are finite. This is [Proposition 4.12](#) for A_c , and the almost-sure condition $\mathbf{1}^\top \mathbf{x}_0^{\ell,f} \neq 1$ for B_f .

Fix $\mathbf{x}_0$ as in [Proposition 4.12](#) which also satisfies the almost-sure conditions of [Theorem 4.10](#) and $\mathbf{1}^\top \mathbf{x}_0^{\ell,f} \neq 1$ for all ℓ ; each of these is a μ_0 -null exception (Step 5 of the proof of [Theorem 4.7](#)), and there are finitely many.

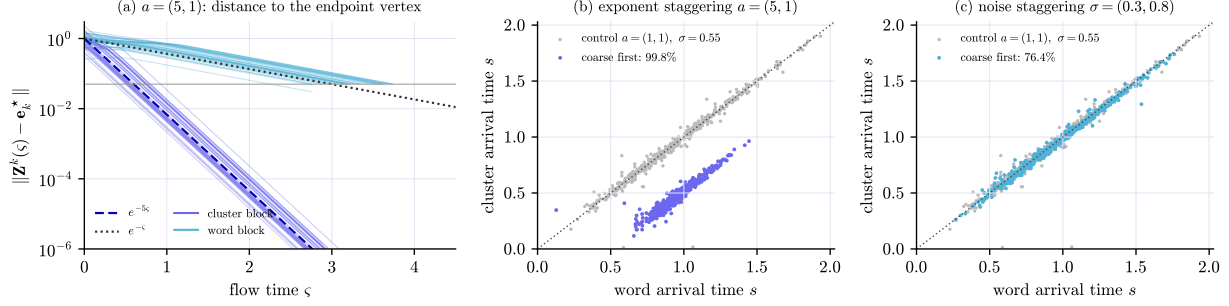


Figure 7 The flow removes the noise of each level on its own schedule ($\Delta^2 \times \Delta^2$, exact field). (a) Distance of the cluster and word blocks to their endpoint vertex along 40 trajectories against the flow time ς , with the slopes $e^{-a_k\varsigma}$ of Theorem 4.7; the grey line is $\varepsilon = 0.05$. (b, c) Arrival times within ε of the vertex in the original time s for 500 prior samples, cluster against word, with the unstaggered control in grey. Exponent staggering separates the arrivals; noise staggering orders them only weakly, as Theorem 4.10 predicts.

Fix ℓ and write $A_c := \|x_0^{\ell,c} - e_{\ell,c}^*\| + K_{\ell,c}$, which is finite by Proposition 4.12, and $B_f := |1^\top x_0^{\ell,f} - 1|/\sqrt{v_f} > 0$. In the notation of the proof of Theorem 4.10, $U_c = \frac{1}{a_c} \ln \frac{A_c}{\varepsilon}$ and $L_f = \frac{1}{a_f} \ln \frac{B_f}{\varepsilon}$. Subtracting $-\frac{\ln \varepsilon}{a_c} + \frac{\ln \varepsilon}{a_f} = -(\frac{1}{a_f} - \frac{1}{a_c}) \ln \frac{1}{\varepsilon}$ from both sides shows that $U_c < L_f$ is equivalent to

$$\left(\frac{1}{a_f} - \frac{1}{a_c}\right) \ln \frac{1}{\varepsilon} > R_\ell := \frac{\ln A_c}{a_c} - \frac{\ln B_f}{a_f} \in [-\infty, \infty).$$

Here $d := \frac{1}{a_f} - \frac{1}{a_c} > 0$ because $a_c > a_f > 0$, and $R_\ell < \infty$ because $A_c < \infty$ and $B_f > 0$. Let $\varepsilon_\ell := \min\{B_f, e^{-\max\{R_\ell, 0\}/d}\} > 0$. For $\varepsilon \in (0, \varepsilon_\ell)$ we have $d \ln \frac{1}{\varepsilon} > \max\{R_\ell, 0\} \geq R_\ell$, so $U_c < L_f$, and $\varepsilon < B_f$, so $L_f > 0$. By the order part of the statement of Theorem 4.10, applicable because $U_c < L_f$ and $L_f > 0$ give $\max\{0, U_c\} < L_f$, block (ℓ, c) comes within ε of its vertex strictly before block (ℓ, f) , in flow time and in the original time s . Take $\varepsilon_0 := \min_\ell \varepsilon_\ell > 0$, a minimum over the finitely many positions. Finally, $K_{\ell,c} < \infty$ almost surely by Proposition 4.12; the events $\{K_{\ell,c} > \bar{K}\}$ decrease as $\bar{K} \uparrow \infty$ with intersection $\{K_{\ell,c} = \infty\}$, a null set, so $\mathbb{P}(K_{\ell,c} > \bar{K}) \rightarrow 0$ by continuity of probability from above. $\square$

Numerical check. On $\Delta^2 \times \Delta^2$ with uniform data on the nine vertex pairs we evaluate the exact field Eq. (5) with M and $\bar{p}^{\ell,k}$ from Eq. (6) (time integral on a logarithmic grid in $1 - t \in [10^{-9}, 1]$, squared distances evaluated directly to avoid cancellation). (a) Integrating in the original time s (RK4, step limited by $0.1/(a_{\max}M)$), the identity $\ln(h^c/h_0^c)/a_c = \ln(h^f/h_0^f)/a_f$ implied by Eq. (16) holds to within $6 \cdot 10^{-5}$ on 24 trajectories for $a = (5, 1)$, $\sigma = 0.55$, down to $\lambda_f = 10^{-3}$, and no coordinate order changes (Theorem 4.9). (b) Integrating $dZ/d\varsigma = b/M$ (RK4, $\Delta\varsigma = 0.02$) from 500 prior samples per setting, with arrival radius $\varepsilon = 0.05$:

setting	coarse first	endpoint = $\arg \max x_0$	median $(a_c \varsigma_c^\varepsilon, a_f \varsigma_f^\varepsilon)$	median / 99% of (K_c, K_f)
$a = (5, 1), \sigma_k = 0.55$	99.8%	100%	(3.20, 3.26)	(0.55, 0.53) / (8.6, 1.8)
$a = (1, 1), \sigma = (0.3, 0.8)$	76.4%	100%	(3.06, 3.42)	(0.27, 0.76) / (2.0, 5.3)
$a = (1, 1), \sigma_k = 0.55$ (control)	47.6%	100%	(3.26, 3.22)	(0.54, 0.55) / (4.3, 4.9)

The integrals K change by less than $5 \cdot 10^{-6}$ (relative) between the cut-offs $\lambda_k = 10^{-4}$ and 10^{-6} , so they are finite on every trajectory; the two-sided bound Eq. (19) holds at every recorded step of 120 stored trajectories (largest ratio of distance to upper bound 0.999), and $a_k \varsigma_k^\varepsilon \approx \ln(1/\varepsilon) + O(1) = 3.0 + O(1)$ for every level, as Eq. (20) predicts. In the original time s the median arrival times are (0.47, 0.95) for exponent staggering, (0.92, 0.94) for noise staggering and (0.95, 0.95) for the control (Fig. 7).

C.11 Proofs for correlated data

Every statement below is reduced to [Theorem 4.15](#), the threshold theorem with positive weights on the competitors, whose proof repeats that of [Theorem 4.2](#). [Theorem 4.16](#) shows that for any data the posterior of a block has this weighted form, with weights determined by the other blocks and hence independent of the block's own noise. [Proposition 4.17](#) bounds the weights when the support of the block does not depend on the context, and [Proposition 4.18](#) and [Theorem 4.19](#) handle context-dependent supports by conditioning on the context. [Corollary 4.20](#) uses the first route for the cluster and the second for the word of a two-level code.

Throughout this subsection $t \in (0, 1)$ is fixed, so every signal-to-noise ratio $\rho_{B'} = (1 - \alpha_{B'})/(\alpha_{B'}\sigma_{B'})$ is a deterministic nonnegative number. Write $\mathbf{j}^*$ for the true configuration ($\mathbf{x}_1 = \mathbf{e}_{\mathbf{j}^*}$). Since $\mathbb{P}(\pi_{\mathbf{j}^*} = 0) = \sum_{\mathbf{j}: \pi_{\mathbf{j}}=0} \pi_{\mathbf{j}} = 0$, we work on the almost sure event $\{\pi_{\mathbf{j}^*} > 0\}$. On it, the true symbol i^* of B satisfies $\pi_B(i^*) \geq \pi_{\mathbf{j}^*} > 0$, and the true context $o^* := \mathbf{j}_{-B}^*$ satisfies $\pi(i^*, o^*) > 0$, so $o^* \in \mathcal{O}^* \neq \emptyset$.

For the conditional statements we use the following standard fact.

Lemma C.10 (Freezing). *Let $\mathbf{g}$ be a random vector in $\mathbb{R}^n$ independent of a σ -algebra $\mathcal{F}$, let θ be an $\mathcal{F}$ -measurable random element of a measurable space $(E, \mathcal{E})$, and let $H : \mathbb{R}^n \times E \rightarrow [0, 1]$ be jointly measurable. Then $h(\vartheta) := \mathbb{E}H(\mathbf{g}, \vartheta)$ is measurable and $\mathbb{E}[H(\mathbf{g}, \theta) \mid \mathcal{F}] = h(\theta)$ almost surely.*

Proof. Measurability of h is part of Tonelli's theorem. Let $F \in \mathcal{F}$. The pair $(\theta, \mathbb{K}_F)$ is $\mathcal{F}$ -measurable, hence independent of $\mathbf{g}$, so the law of $(\mathbf{g}, (\theta, \mathbb{K}_F))$ is the product $\mathbb{P}_{\mathbf{g}} \otimes \mathbb{P}_{(\theta, \mathbb{K}_F)}$. By Fubini's theorem,

$$\mathbb{E}[H(\mathbf{g}, \theta) \mathbb{K}_F] = \int \left(\int H(\mathbf{z}, \vartheta) \mathbb{P}_{\mathbf{g}}(d\mathbf{z}) \right) a \mathbb{P}_{(\theta, \mathbb{K}_F)}(d\vartheta, da) = \mathbb{E}[h(\theta) \mathbb{K}_F].$$

As $h(\theta)$ is $\mathcal{F}$ -measurable, this is the defining property of $\mathbb{E}[H(\mathbf{g}, \theta) \mid \mathcal{F}]$. □

Proof of [Theorem 4.15](#)

Proof. Idea. The weights enter Σ_w linearly, so the proof of [Theorem 4.2](#) goes through with two changes: in (a) the mean of the truncated sum acquires the factor $\bar{w}$, and in (b) the single dominant term carries a weight of at least $\hat{w}$. The conditional version follows by applying the unconditional statement at frozen values of the index sets and weights, which is legitimate because they are independent of $\mathbf{g}$ ([Lemma C.10](#)).

Write $g_\star := g_{i^*}$ and $s := s_{m+1} = \sqrt{2 \log m}$. Since $e^{\rho(g_j - g_\star) - \rho^2} = e^{-\rho g_\star - \rho^2} e^{\rho g_j}$,

$$\Sigma_w = e^{-\rho g_\star - \rho^2} S^w, \quad S^w := \sum_{j \in J} w_j e^{\rho g_j}, \quad (74)$$

and since $\Sigma_w \geq 0$,

$$1 - \pi_\star = \frac{\Sigma_w}{1 + \Sigma_w} \leq \Sigma_w. \quad (75)$$

(a) Fix $u, u' > 0$ and $\rho \geq s + 2u$; let $M := s + u$. Define

$$A = \{g_\star \geq -u\}, \quad B = \{\max_{j \in J} g_j \leq M\}, \quad C = \{S_M^w \leq e^{u'} \mathbb{E} S_M^w\}, \quad S_M^w := \sum_{j \in J} w_j e^{\rho g_j} \mathbb{K}\{g_j \leq M\}.$$

Probabilities. By symmetry of $\mathcal{N}(0, 1)$ and [Lemma C.3](#), $\mathbb{P}(A^c) = \mathbb{P}(-g_\star > u) = Q(u) \leq e^{-u^2/2}$. Since $i^* \notin J$, the variables $(g_j)_{j \in J}$ are $m \geq 1$ independent standard normals, so [Lemma C.4](#) with $s = \sqrt{2 \log m}$ gives $\mathbb{P}(B^c) \leq e^{-u^2/2}$. Each term of S_M^w lies in $[0, w_j e^{\rho M}]$, so $0 \leq S_M^w \leq m \bar{w} e^{\rho M}$, and $\mathbb{E} S_M^w > 0$ by [Eq. \(76\)](#) below (all $w_j > 0$ and $\Phi > 0$). Markov's inequality gives $\mathbb{P}(C^c) = \mathbb{P}(S_M^w > e^{u'} \mathbb{E} S_M^w) \leq e^{-u'}$. By the union bound over A^c, B^c, C^c , $\mathbb{P}(A \cap B \cap C) \geq 1 - 2e^{-u^2/2} - e^{-u'}$.

The truncated sum. By linearity, [Lemma C.5](#) applied to each g_j , and $\sum_{j \in J} w_j = m \bar{w}$,

$$\mathbb{E} S_M^w = \sum_{j \in J} w_j \mathbb{E}[e^{\rho g_j} \mathbb{K}\{g_j \leq M\}] = \sum_{j \in J} w_j e^{\rho^2/2} \Phi(M - \rho) = m \bar{w} e^{\rho^2/2} \Phi(M - \rho). \quad (76)$$

As $\rho \geq s + 2u > M$, $\Phi(M - \rho) = Q(\rho - M)$ with $\rho - M > 0$, and [Lemma C.3](#) gives $\Phi(M - \rho) \leq e^{-(\rho - M)^2/2}$. Hence

$$\mathbb{E}S_M^w \leq m\bar{w} \exp\left(\frac{\rho^2}{2} - \frac{(\rho - M)^2}{2}\right) = m\bar{w} \exp\left(\rho M - \frac{M^2}{2}\right),$$

which is [Eq. \(63\)](#) with the extra factor $\bar{w}$.

The bound on the good event. On B , $\mathbb{P}\{g_j \leq M\} = 1$ for all $j \in J$, so $S^w = S_M^w$; on C , $S_M^w \leq e^{u'} \mathbb{E}S_M^w$; on A , $-\rho g_\star \leq \rho u$ because $\rho \geq 0$. By [Eq. \(74\)](#), on $A \cap B \cap C$,

$$\Sigma_w = e^{-\rho g_\star - \rho^2} S_M^w \leq e^{\rho u - \rho^2} e^{u'} m\bar{w} e^{\rho M - M^2/2} = \bar{w} \exp\left(u' + \rho u - \rho^2 + \rho M + \left(\log m - \frac{M^2}{2}\right)\right).$$

Since $\log m = s^2/2$ and $M^2 = s^2 + 2su + u^2$, $\log m - M^2/2 = -su - u^2/2 \leq 0$. Dropping this term and substituting $M = s + u$,

$$\Sigma_w \leq \bar{w} \exp(u' + \rho u - \rho^2 + \rho(s + u)) = \bar{w} \exp(u' - \rho(\rho - s - 2u)).$$

Together with [Eq. \(75\)](#), the chain $1 - \pi_\star \leq \Sigma_w \leq \bar{w} \exp(u' - \rho(\rho - s - 2u))$ holds on $A \cap B \cap C$.

(b) Let $s' := s_{m'+1} = \sqrt{2 \log m'}$, which is positive because $m' \geq 2$, and let $G' := \max_{j \in J'} g_j$ and $\hat{j} \in J'$ an index attaining it. All terms of Σ_w are nonnegative and $\hat{j} \in J' \subseteq J$, so with $z := \rho(G' - g_\star) - \rho^2$,

$$\Sigma_w \geq w_{\hat{j}} e^{\rho(g_{\hat{j}} - g_\star) - \rho^2} = w_{\hat{j}} e^z \geq \hat{w} e^z, \quad \pi_\star = \frac{1}{1 + \Sigma_w} \leq \frac{1}{1 + \hat{w} e^z} \leq \frac{e^{-z}}{\hat{w}}.$$

Let $D = \{g_\star \leq u\}$ and $E = \{G' \geq s' - h\}$. By [Lemma C.3](#), $\mathbb{P}(D^c) = Q(u) \leq e^{-u^2/2}$. The variables $(g_j)_{j \in J'}$ are m' independent standard normals, so [Lemmas C.6](#) and [C.7](#) with $x := s' - h > 0$ give

$$\mathbb{P}(E^c) = \mathbb{P}(G' < x) \leq \exp(-m' Q(x)) \leq \exp\left(-m' \frac{x}{1+x^2} \varphi(x)\right).$$

Using $m' e^{-s'^2/2} = m' e^{-\log m'} = 1$, $m' \varphi(x) = \frac{m'}{\sqrt{2\pi}} e^{-s'^2/2} e^{s'h - h^2/2} = \frac{1}{\sqrt{2\pi}} e^{s'h - h^2/2}$, so $m' \frac{x}{1+x^2} \varphi(x) = \Lambda_{m'+1}(h)$ (the function of [Eq. \(13\)](#) with $n = m' + 1$, $s_n = s'$) and $\mathbb{P}(E^c) \leq e^{-\Lambda_{m'+1}(h)}$. By the union bound, $\mathbb{P}(D \cap E) \geq 1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$. On $D \cap E$, $G' - g_\star \geq s' - h - u$, so, as $\rho \geq 0$ and $\rho \leq s' - h - u - \eta$,

$$z \geq \rho(s' - h - u) - \rho^2 = \rho(s' - h - u - \rho) \geq \rho\eta,$$

and $\pi_\star \leq e^{-z}/\hat{w} \leq e^{-\rho\eta}/\hat{w}$.

Conditional version. We allow $i^\star$, J , J' and the weights to be $\mathcal{F}$ -measurable. Collect them in $\theta := (i^\star, J, J', (w_j)_{j \in J})$, an $\mathcal{F}$ -measurable random element of the space E of such tuples, a finite disjoint union (over the choices of $(i^\star, J, J')$) of open orthants with their Borel σ -algebras. For fixed $(i^\star, J, J')$, the quantities Σ_w , $\pi_\star$, $\bar{w}$ and $\hat{w}$ are continuous in $(\mathbf{g}, (w_j)_j)$, so the indicator $H_a(\mathbf{g}, \vartheta)$ of the event in (a) and the indicator $H_b(\mathbf{g}, \vartheta)$ of the event in (b), evaluated at the parameter $\vartheta \in E$, are jointly Borel. For deterministic ϑ satisfying the hypotheses of (a) (resp. (b)), the statements proved above say $\mathbb{E}H_a(\mathbf{g}, \vartheta) \geq 1 - 2e^{-u^2/2} - e^{-u'}$ (resp. $\mathbb{E}H_b(\mathbf{g}, \vartheta) \geq 1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$), because $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. By [Lemma C.10](#), almost surely on the $\mathcal{F}$ -measurable event that θ satisfies these hypotheses, $\mathbb{P}(\text{event of (a)} \mid \mathcal{F}) = \mathbb{E}[H_a(\mathbf{g}, \vartheta)]_{\vartheta=\theta} \geq 1 - 2e^{-u^2/2} - e^{-u'}$, and likewise for (b). $\square$

Proof of [Theorem 4.16](#)

Proof. Idea. Without the product structure the posterior no longer factorises, but the likelihood still does, into one factor $\kappa_{B''}(\mathbf{x}^{B''}) e^{\beta_{B''} x_{j_{B''}}^{B''}}$ per block. The joint posterior of (i, o) is therefore proportional to $\pi(i, o) e^{\beta x_i^{\ell, k}} L(o)$, and summing out the context leaves the single-block softmax with an extra factor $N_i = \sum_o \pi(i, o) L(o)$, which becomes the offset Δ_j . It then remains to check that the offsets depend only on $\mathbf{x}_1$ and on the noise of the other blocks, which is independent of $\mathbf{g}$ because the prior covariance is block diagonal.

For a block B'' write $n_{B''}$, $\alpha_{B''}$, $\sigma_{B''}$, $\beta_{B''}$ for its size, schedule value, prior scale and inverse temperature, and $p^{B''}(\cdot \mid i)$ for the density of $\mathcal{N}((1 - \alpha_{B''})\mathbf{e}_i, \alpha_{B''}^2 \sigma_{B''}^2 \mathbf{I})$. By [Lemma B.1](#), $p(\mathbf{x} \mid t, \mathbf{j}) = \prod_{B''} p^{B''}(\mathbf{x}^{B''} \mid j_{B''})$.

Expanding the square as in Step 7 of the proof of [Proposition 4.1](#), $\|\mathbf{y} - (1 - \alpha_{B''})\mathbf{e}_i\|^2 = \|\mathbf{y}\|^2 - 2(1 - \alpha_{B''})y_i + (1 - \alpha_{B''})^2$, so

$$p^{B''}(\mathbf{y} | i) = \kappa_{B''}(\mathbf{y}) e^{\beta_{B''} y_i}, \quad \kappa_{B''}(\mathbf{y}) := (2\pi\alpha_{B''}^2\sigma_{B''}^2)^{-n_{B''}/2} \exp\left(-\frac{\|\mathbf{y}\|^2 + (1 - \alpha_{B''})^2}{2\alpha_{B''}^2\sigma_{B''}^2}\right) > 0,$$

with $\kappa_{B''}$ independent of i . For $\mathbf{j} = (i, o)$, separating the factor of B from those of the blocks $B' \neq B$,

$$\pi_{\mathbf{j}} p(\mathbf{x} | t, \mathbf{j}) = \pi(i, o) \prod_{B''} \kappa_{B''}(\mathbf{x}^{B''}) e^{\beta_{B''} x_{\mathbf{j}B''}^{B''}} = \kappa(\mathbf{x}) \pi(i, o) e^{\beta x_i^{\ell,k}} L(o), \quad \kappa(\mathbf{x}) := \prod_{B''} \kappa_{B''}(\mathbf{x}^{B''}) > 0. \quad (77)$$

Bayes' rule at fixed t ([Eq. \(54\)](#), whose derivation does not use the product structure) and summation over the contexts give, with $N_i := \sum_o \pi(i, o) L(o)$,

$$\mathbb{P}(j_{\ell,k} = j | \mathbf{I}_t = \mathbf{x}) = \frac{\sum_o \kappa(\mathbf{x}) \pi(j, o) e^{\beta x_j^{\ell,k}} L(o)}{\sum_i \sum_o \kappa(\mathbf{x}) \pi(i, o) e^{\beta x_i^{\ell,k}} L(o)} = \frac{e^{\beta x_j^{\ell,k}} N_j}{\sum_i e^{\beta x_i^{\ell,k}} N_i}. \quad (78)$$

As $L(o) > 0$ for all o , $N_j > 0$ iff $\pi(j, o) > 0$ for some o , i.e. iff $j \in S_B$. Hence the probability vanishes for $j \notin S_B$, while $N_{i^*} > 0$ because $i^* \in S_B$. For $j \in S_B$ both probabilities in [Eq. \(23\)](#) are positive and

$$\log \frac{\mathbb{P}(j_{\ell,k} = j | \mathbf{I}_t)}{\mathbb{P}(j_{\ell,k} = i^* | \mathbf{I}_t)} = \beta(x_j^{\ell,k} - x_{i^*}^{\ell,k}) + \log \frac{N_j}{N_{i^*}} = \beta(x_j^{\ell,k} - x_{i^*}^{\ell,k}) + \Delta_j.$$

Since $\mathbf{x}^{\ell,k} = (1 - \alpha)\mathbf{e}_{i^*} + \alpha\sigma\mathbf{g}$, i.e. $x_i^{\ell,k} = (1 - \alpha)\mathbb{1}\{i = i^*\} + \alpha\sigma g_i$, and $\beta\alpha\sigma = (1 - \alpha)/(\alpha\sigma) = \rho$, $\beta(1 - \alpha) = (1 - \alpha)^2/(\alpha^2\sigma^2) = \rho^2$, for $j \neq i^*$

$$\beta(x_j^{\ell,k} - x_{i^*}^{\ell,k}) = \beta\alpha\sigma(g_j - g_{i^*}) - \beta(1 - \alpha) = \rho(g_j - g_{i^*}) - \rho^2, \quad (79)$$

and for $j = i^*$ both sides of [Eq. \(23\)](#) vanish ($\Delta_{i^*} = 0$). This proves [Eq. \(23\)](#). As the block posterior sums to one over S_B ,

$$\frac{1}{\pi_{i^*}^{\ell,k}} = \sum_{j \in S_B} \frac{\mathbb{P}(j_{\ell,k} = j | \mathbf{I}_t)}{\mathbb{P}(j_{\ell,k} = i^* | \mathbf{I}_t)} = 1 + \sum_{j \in S_B \setminus \{i^*\}} e^{\Delta_j} e^{\rho(g_j - g_{i^*}) - \rho^2},$$

which is [Eq. \(22\)](#) with $J = S_B \setminus \{i^*\}$ and $w_j = e^{\Delta_j} = N_j/N_{i^*} \in (0, \infty)$. (If $J = \emptyset$, then $\pi_{i^*}^{\ell,k} = 1$ and there is nothing to prove.)

Measurability and independence. For $B' \neq B$, $\mathbf{x}^{B'} = \mathbf{I}_t^{B'} = (1 - \alpha_{B'})\mathbf{x}_1^{B'} + \alpha_{B'}\mathbf{x}_0^{B'}$ is $\mathcal{F}_{-B}$ -measurable, hence so are $L(o)$ for every o , all N_i , and the Δ_j ; i^* is a function of $\mathbf{x}_1^{\ell,k}$, so $J = S_B \setminus \{i^*\}$ is $\mathcal{F}_{-B}$ -measurable as well. The prior $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \Sigma)$ is Gaussian with block-diagonal covariance, so $\mathbf{x}_0^{\ell,k}$ and $(\mathbf{x}_0^{B'})_{B' \neq B}$ are uncorrelated jointly Gaussian, hence independent; and $\mathbf{x}_0$ is independent of $\mathbf{x}_1$. For Borel sets U, U', U'' therefore

$$\mathbb{P}(\mathbf{x}_0^{\ell,k} \in U, (\mathbf{x}_0^{B'})_{B' \neq B} \in U', \mathbf{x}_1 \in U'') = \mathbb{P}(\mathbf{x}_0^{\ell,k} \in U) \mathbb{P}((\mathbf{x}_0^{B'})_{B' \neq B} \in U') \mathbb{P}(\mathbf{x}_1 \in U''),$$

so $\mathbf{g} = \mathbf{x}_0^{\ell,k}/\sigma$ is independent of $\mathcal{F}_{-B}$. Since $\pi_{i^*}^{\ell,k}$ has the form [Eq. \(22\)](#) with $\mathcal{F}_{-B}$ -measurable $(i^*, J, (w_j))$, the conditional version of [Theorem 4.15](#) applies with $\mathcal{F} = \mathcal{F}_{-B}$. $\square$

Proof of [Proposition 4.17](#)

Proof. Idea. The offset compares N_j with N_{i^*} , two sums over contexts with the same evidence weights and different prior weights, and a ratio of two such sums is at most the largest ratio of corresponding terms. For $F = \emptyset$ this gives Γ_B directly. For general F we first replace the evidence of the blocks in F by its maximum or its minimum over their symbols, at the cost of the factor e^{D_F} , and then compare the marginal weights $\pi(i, r)$. The tail bound on the evidence range is [Lemma C.4](#) applied to the maximum and to the minimum of the block noise.

Write $N_i = \sum_o \pi(i, o) L(o)$ as in the proof of [Theorem 4.16](#), so $e^{\Delta_j} = N_j/N_{i^*}$. Splitting the product defining $L(o)$ into the blocks of F and the remaining blocks $B' \notin F \cup \{B\}$, $L(o) = L_F(f) L_r(r)$ with

$L_F(f) := \prod_{B' \in F} e^{\beta_{B'} x_{f_{B'}}^{B'}}$ and L_r the product over the remaining blocks; both are positive. Let $\bar{L}_F := \max_f L_F(f)$ and $\underline{L}_F := \min_f L_F(f)$ over all $f \in \prod_{B' \in F} [v_{B'}]$. Since $L_F(f) = \exp(\sum_{B' \in F} \beta_{B'} x_{f_{B'}}^{B'})$, $\exp$ is increasing, and each summand depends only on its own coordinate $f_{B'}$ of the product set,

$$\bar{L}_F = \exp\left(\sum_{B' \in F} \max_i \beta_{B'} x_i^{B'}\right), \quad \underline{L}_F = \exp\left(\sum_{B' \in F} \min_i \beta_{B'} x_i^{B'}\right), \quad \frac{\bar{L}_F}{\underline{L}_F} = e^{D_F}.$$

For every i , since $\pi \geq 0$ and $\underline{L}_F \leq L_F(f) \leq \bar{L}_F$,

$$\underline{L}_F \sum_r L_r(r) \pi(i, r) \leq N_i = \sum_r L_r(r) \sum_f \pi(i, f, r) L_F(f) \leq \bar{L}_F \sum_r L_r(r) \pi(i, r). \quad (80)$$

Let $\mathcal{R} := \{r : \sum_i \pi(i, r) > 0\}$. For $r \notin \mathcal{R}$, $\pi(i, r) = 0$ for all i , so the sums in Eq. (80) may be restricted to $\mathcal{R}$. Fix $j \in S_B$. For $r \in \mathcal{R}$, rectangular support gives $\pi(i^*, r) > 0$ and $\pi(j, r) > 0$ (both $i^*, j \in S_B$), and the definition of $\Gamma_{B|F}$ (a maximum over $r \in \mathcal{R}$) gives $\pi(j, r) \leq \Gamma_{B|F} \pi(i^*, r)$. The set $\mathcal{R}$ is nonempty since it contains the true r . Hence, by the upper bound in Eq. (80) for $i = j$ and the lower bound for $i = i^*$,

$$e^{\Delta_j} = \frac{N_j}{N_{i^*}} \leq \frac{\bar{L}_F \sum_{r \in \mathcal{R}} L_r(r) \pi(j, r)}{\underline{L}_F \sum_{r \in \mathcal{R}} L_r(r) \pi(i^*, r)} \leq \frac{\bar{L}_F}{\underline{L}_F} \cdot \frac{\Gamma_{B|F} \sum_{r \in \mathcal{R}} L_r(r) \pi(i^*, r)}{\sum_{r \in \mathcal{R}} L_r(r) \pi(i^*, r)} = e^{D_F} \Gamma_{B|F},$$

the denominators being positive. The same computation with j and i^* exchanged (it uses only $j, i^* \in S_B$) gives $e^{-\Delta_j} = N_{i^*}/N_j \leq e^{D_F} \Gamma_{B|F}$. Taking logarithms, $-\log \Gamma_{B|F} - D_F \leq \Delta_j \leq \log \Gamma_{B|F} + D_F$. For $F = \emptyset$, f is the empty tuple, $r = o$, $L_F \equiv 1$, $D_\emptyset = 0$ (empty sum), and for o of positive mass $\pi(i, o)/\pi(i', o) = \pi(i | o)/\pi(i' | o)$ (divide numerator and denominator by $\sum_{i''} \pi(i'', o)$).

Evidence range. Let B' have v' vertices, true symbol i' and $\mathbf{g}' := \mathbf{x}_0^{B'}/\sigma_{B'} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{v'})$. Then $x_i^{B'} = (1 - \alpha_{B'}) \mathbb{1}\{i = i'\} + \alpha_{B'} \sigma_{B'} g'_i$, and as in Eq. (79), $\beta_{B'}(1 - \alpha_{B'}) = \rho_{B'}^2$ and $\beta_{B'} \alpha_{B'} \sigma_{B'} = \rho_{B'}$, so $\beta_{B'} x_i^{B'} = \rho_{B'}^2 \mathbb{1}\{i = i'\} + \rho_{B'} g'_i$. For any reals, $\max_i (a_i + b_i) \leq \max_i a_i + \max_i b_i$ and $\min_i (a_i + b_i) \geq \min_i a_i + \min_i b_i$; with $a_i = \rho_{B'}^2 \mathbb{1}\{i = i'\} \in [0, \rho_{B'}^2]$ and $b_i = \rho_{B'} g'_i$ (recall $\rho_{B'} \geq 0$),

$$\max_i \beta_{B'} x_i^{B'} - \min_i \beta_{B'} x_i^{B'} \leq \rho_{B'}^2 + \rho_{B'} (\max_i g'_i - \min_i g'_i).$$

The vectors $\mathbf{g}'$ and $-\mathbf{g}'$ both have i.i.d. standard normal coordinates, so Lemma C.4 with $m = v'$ gives $\mathbb{P}(\max_i g'_i > \sqrt{2 \log v'} + u) \leq e^{-u^2/2}$ and $\mathbb{P}(-\min_i g'_i > \sqrt{2 \log v'} + u) \leq e^{-u^2/2}$. By the union bound, with probability at least $1 - 2e^{-u^2/2}$ neither event occurs, and then $\max_i g'_i - \min_i g'_i \leq 2(\sqrt{2 \log v'} + u)$, which together with the previous display gives the claim. $\square$

Proof of Proposition 4.18

Proof. Idea. The identity is the law of total probability over the context. Given the context o , block B sees only the prior weights $\pi(\cdot | o)$ and its own evidence, so its conditional posterior is a single-block expression with weights $w_j(o)$. The two inequalities follow because $\pi_\star^{\ell, k}$ is an average of the $\pi_\star(o)$ with weights $\mathbb{P}(o | \mathbf{I}_t)$ that sum to one.

Let $\mathbf{x} = \mathbf{I}_t$. By Eq. (77) and Bayes' rule, $\mathbb{P}(j_{\ell, k} = i, o | \mathbf{I}_t = \mathbf{x}) = \kappa(\mathbf{x}) \pi(i, o) e^{\beta x_i^{\ell, k}} L(o) / Z(\mathbf{x})$, so the posterior of the context is

$$\mathbb{P}(o | \mathbf{I}_t = \mathbf{x}) = \sum_i \mathbb{P}(j_{\ell, k} = i, o | \mathbf{I}_t = \mathbf{x}) = \frac{\kappa(\mathbf{x}) L(o)}{Z(\mathbf{x})} \sum_i \pi(i, o) e^{\beta x_i^{\ell, k}},$$

which is positive iff $\sum_i \pi(i, o) > 0$, and $\sum_o \mathbb{P}(o | \mathbf{I}_t) = 1$. By the law of total probability over the context, $\pi_\star^{\ell, k} = \sum_o \mathbb{P}(o | \mathbf{I}_t) \mathbb{P}(j_{\ell, k} = i^* | o, \mathbf{I}_t)$, the sum running over contexts with $\mathbb{P}(o | \mathbf{I}_t) > 0$. For such o , dividing the joint posterior by the context posterior, cancelling $\kappa(\mathbf{x}) L(o) / Z(\mathbf{x})$, and then dividing numerator and denominator by $\sum_{i''} \pi(i'', o) > 0$,

$$\mathbb{P}(j_{\ell, k} = i | o, \mathbf{I}_t = \mathbf{x}) = \frac{\pi(i, o) e^{\beta x_i^{\ell, k}}}{\sum_{i'} \pi(i', o) e^{\beta x_{i'}^{\ell, k}}} = \frac{\pi(i | o) e^{\beta x_i^{\ell, k}}}{\sum_{i' \in S(o)} \pi(i' | o) e^{\beta x_{i'}^{\ell, k}}},$$

where the terms with $i' \notin S(o)$ vanish. If $o \notin \mathcal{O}^*$, then $\pi(i^* | o) = 0$ and this vanishes for $i = i^*$. If $o \in \mathcal{O}^*$, dividing numerator and denominator by $\pi(i^* | o)e^{\beta x_{i^*}^{\ell,k}} > 0$ and using Eq. (79),

$$\mathbb{P}(j_{\ell,k} = i^* | o, \mathbf{I}_t) = \left(1 + \sum_{j \in S(o) \setminus \{i^*\}} \frac{\pi(j | o)}{\pi(i^* | o)} e^{\beta(x_j^{\ell,k} - x_{i^*}^{\ell,k})}\right)^{-1} = \pi_*(o).$$

This proves the identity.

Let $\mathcal{G} \subseteq \mathcal{O}^*$. Using $\sum_o \mathbb{P}(o | \mathbf{I}_t) = 1$, the identity, and $\mathcal{G} \subseteq \mathcal{O}^*$,

$$\begin{aligned} 1 - \pi_*^{\ell,k} &= \sum_{o \notin \mathcal{G}} \mathbb{P}(o | \mathbf{I}_t) (1 - \mathbb{P}\{o \in \mathcal{O}^*\} \pi_*(o)) + \sum_{o \in \mathcal{G}} \mathbb{P}(o | \mathbf{I}_t) (1 - \pi_*(o)) \\ &\leq \mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t) + \sup_{o \in \mathcal{G}} (1 - \pi_*(o)) \sum_{o \in \mathcal{G}} \mathbb{P}(o | \mathbf{I}_t) \leq \mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t) + \sup_{o \in \mathcal{G}} (1 - \pi_*(o)), \end{aligned}$$

where the first inequality uses $0 \leq \pi_*(o) \leq 1$ and the second $\sum_{o \in \mathcal{G}} \mathbb{P}(o | \mathbf{I}_t) \leq 1$. Finally, since $\pi_*(o) \geq 0$, $\pi_*^{\ell,k} = \sum_{o \in \mathcal{O}^*} \mathbb{P}(o | \mathbf{I}_t) \pi_*(o) \leq \sup_{o \in \mathcal{O}^*} \pi_*(o) \sum_{o \in \mathcal{O}^*} \mathbb{P}(o | \mathbf{I}_t) \leq \sup_{o \in \mathcal{O}^*} \pi_*(o)$. $\square$

Proof of Theorem 4.19

Proof. Idea. By Proposition 4.18 it suffices to bound $1 - \pi_*(o)$ uniformly over $o \in \mathcal{G}$ for (a), and $\pi_*(o)$ uniformly over $o \in \mathcal{O}^*$ for (b). Each $\pi_*(o)$ is a weighted threshold expression, so we compare it with a single expression that does not depend on o : for (a) we enlarge the competitor set to $S_{\mathcal{G}}$ and raise all weights to $\Gamma_{\mathcal{G}}$, for (b) we shrink it to J' and lower all weights to $1/\Gamma'$. Theorem 4.15 applies to these expressions conditionally on $\mathbf{x}_1$, because given $\mathbf{x}_1$ the sets and weights are fixed and $\mathbf{g}$ is still standard Gaussian.

Condition on $\mathbf{x}_1$. Then i^* , $\mathcal{O}^*$, the sets $S(o)$ and the weights $w_j(o)$ are fixed (they depend only on μ_1 and i^*), as are $\mathcal{G}$, $S_{\mathcal{G}}$, J' , $\Gamma_{\mathcal{G}}$ and Γ' (we take $\mathcal{G}$ $\sigma(\mathbf{x}_1)$ -measurable). As $\mathbf{x}_0$ is independent of $\mathbf{x}_1$, $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ conditionally on $\mathbf{x}_1$, so Theorem 4.15 applies conditionally on $\mathcal{F} = \sigma(\mathbf{x}_1)$.

(a) Let $o \in \mathcal{G}$. Then $S(o) \setminus \{i^*\} \subseteq S_{\mathcal{G}}$ and $w_j(o) \leq \Gamma_{\mathcal{G}}$ for $j \in S(o) \setminus \{i^*\}$. By Eq. (75) applied to $\pi_*(o)$, and since all terms are nonnegative,

$$1 - \pi_*(o) \leq \sum_{j \in S(o) \setminus \{i^*\}} w_j(o) e^{\rho(g_j - g_{i^*}) - \rho^2} \leq \Gamma_{\mathcal{G}} \sum_{j \in S_{\mathcal{G}}} e^{\rho(g_j - g_{i^*}) - \rho^2} =: \Gamma_{\mathcal{G}} \Sigma_{\mathcal{G}}, \quad (81)$$

a bound that does not depend on o . $\Sigma_{\mathcal{G}}$ is the sum Eq. (22) with $J = S_{\mathcal{G}} \subseteq [n] \setminus \{i^*\}$, $|J| = m \geq 1$ and unit weights ($\bar{w} = 1$). By Theorem 4.15(a), conditionally on $\mathbf{x}_1$, the event $\{\Sigma_{\mathcal{G}} \leq \exp(u' - \rho(\rho - s_{m+1} - 2u))\}$ has probability at least $1 - 2e^{-u^2/2} - e^{-u'}$. On this event, Eq. (81) gives $\sup_{o \in \mathcal{G}} (1 - \pi_*(o)) \leq \Gamma_{\mathcal{G}} \exp(u' - \rho(\rho - s_{m+1} - 2u))$, and the first inequality of Proposition 4.18 gives the claim. The context term $\mathbb{P}(o \notin \mathcal{G} | \mathbf{I}_t)$ is not estimated: the inequality of Proposition 4.18 holds pathwise.

(b) Let $o \in \mathcal{O}^*$. By definition of J' as an intersection, $J' \subseteq S(o) \setminus \{i^*\}$, and by definition of Γ' , $w_j(o) \geq 1/\Gamma'$ for $j \in J'$ ($\Gamma' \in (0, \infty)$, as the weights are positive and there are finitely many contexts). Dropping the nonnegative terms with $j \notin J'$ and bounding the remaining weights from below,

$$\pi_*(o) \leq \left(1 + \frac{1}{\Gamma'} \sum_{j \in J'} e^{\rho(g_j - g_{i^*}) - \rho^2}\right)^{-1} =: \tilde{\pi}_*,$$

uniformly in $o \in \mathcal{O}^*$. Now $\tilde{\pi}_*$ is the π_* of Theorem 4.15 with $J = J'$ and constant weights $w_j = 1/\Gamma'$. Part (b) of that theorem, applied with J' itself as the subset ($m' \geq 2$, $\hat{w} = 1/\Gamma'$), gives $\tilde{\pi}_* \leq \Gamma' e^{-\rho\eta}$ with conditional probability at least $1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$. On this event, the last inequality of Proposition 4.18 gives $\pi_*^{\ell,k} \leq \sup_{o \in \mathcal{O}^*} \pi_*(o) \leq \tilde{\pi}_* \leq \Gamma' e^{-\rho\eta}$. $\square$

Proof of Corollary 4.20

Proof. Idea. The two levels need the two routes of the preceding results. The cluster block has full support in every context, so (i) is Theorem 4.16 with offsets bounded by Proposition 4.17, taking for F the word block of the same position, which costs its evidence range D_2 and a tail bound for it. The support of the

word block depends on the cluster, so (ii) and (iii) use the mixture over contexts and [Theorem 4.19](#), with $\mathcal{G}$ the contexts that contain the true cluster; the posterior mass outside $\mathcal{G}$ is exactly the cluster error $1 - \pi_\star^{\ell,1}$, which is why the bound for the token contains the bound for its cluster. The first paragraph derives from (H1) and (H2) the support properties that these choices require.

At position ℓ the blocks are the cluster block $(\ell, 1)$ and the word block $(\ell, 2)$; a configuration has positive mass only if $w_{\ell'} \in S(c_{\ell'})$ at every position ℓ' . Two consequences of the hypotheses are used repeatedly. First, $\Gamma_1, \Gamma_2 \geq 1$ (take $c = c'$, resp. $w = w'$). Second, if $\Gamma_1 < \infty$, then for every r of positive mass and every $c' \in [v_1]$, $\mathbb{P}(c_\ell = c' \mid r) > 0$: some c has $\mathbb{P}(c_\ell = c \mid r) > 0$ because these probabilities sum to one, and $\mathbb{P}(c_\ell = c' \mid r) = 0$ would make the ratio for (c, c') infinite. In the same way, (H2) gives

$$\mathbb{P}(c_\ell = c, r) > 0 \implies \{w : \mathbb{P}(c_\ell = c, w_\ell = w, r) > 0\} = S(c), \quad (82)$$

since the conditional law of w_ℓ given $(c_\ell = c, r)$ is carried by $S(c)$ and, by (H2), cannot vanish at a point of $S(c)$.

(i) Take $B = (\ell, 1)$ in [Theorem 4.16](#) and $F = \{(\ell, 2)\}$ in [Proposition 4.17](#). Then f is the word index at position ℓ and r the symbols at all other positions, so $\pi(c, r) = \sum_w \pi(c, w, r) = \mathbb{P}(c_\ell = c, r)$ and, for r of positive mass, $\pi(c, r)/\pi(c', r) = \mathbb{P}(c_\ell = c \mid r)/\mathbb{P}(c_\ell = c' \mid r)$. By the second consequence above, every cluster has positive probability in every context of positive mass; hence $S_B = [v_1]$, rectangular support holds, and $\Gamma_{B|F} = \Gamma_1$. [Proposition 4.17](#) gives $\Delta_c \leq \log \Gamma_1 + D_2$ for all c , with D_2 the evidence range of block $(\ell, 2)$, so $w_c = e^{\Delta_c} \leq \Gamma_1 e^{D_2}$ and $\bar{w} \leq \Gamma_1 e^{D_2}$. By [Theorem 4.16](#), [Theorem 4.15\(a\)](#) applies conditionally on $\mathcal{F}_{-B}$ with $J = [v_1] \setminus \{c^\star\}$, $m = v_1 - 1$ and $s_{m+1} = s_{v_1}$: on an event $\mathcal{E}_a$ with $\mathbb{P}(\mathcal{E}_a \mid \mathcal{F}_{-B}) \leq 2e^{-u^2/2} + e^{-u'}$,

$$1 - \pi_\star^{\ell,1} \leq \bar{w} \exp(u' - \rho_1(\rho_1 - s_{v_1} - 2u)) \leq \Gamma_1 e^{D_2} \exp(u' - \rho_1(\rho_1 - s_{v_1} - 2u)).$$

Taking expectations, $\mathbb{P}(\mathcal{E}_a) \leq 2e^{-u^2/2} + e^{-u'}$. By the last part of [Proposition 4.17](#) with $B' = (\ell, 2)$, $v' = v_2$, $\rho_{B'} = \rho_2$, the event $\mathcal{E}_b := \{D_2 \leq \bar{D}_2\}$ has $\mathbb{P}(\mathcal{E}_b) \leq 2e^{-u^2/2}$. On $\mathcal{E}_a \cap \mathcal{E}_b$ the bound of (i) holds, and by the union bound $\mathbb{P}(\mathcal{E}_a \cap \mathcal{E}_b) \geq 1 - 4e^{-u^2/2} - e^{-u'}$. (Both bounds also hold conditionally on $\mathbf{x}_1$: the first because $\sigma(\mathbf{x}_1) \subseteq \mathcal{F}_{-B}$, the second because $\mathbf{x}_0^{(\ell,2)}$ is independent of $\mathbf{x}_1$.)

(ii) Take $B = (\ell, 2)$, $i^\star = w^\star$, with context $o = (c_\ell, r)$, which contains the cluster of the same position. Let $\mathcal{G} := \{o = (c^\star, r) : \mathbb{P}(c_\ell = c^\star, r) > 0\}$. For $o \in \mathcal{G}$, [Eq. \(82\)](#) gives $S(o) = S(c^\star) \ni w^\star$, so $\pi(w^\star, o) > 0$ and $\mathcal{G} \subseteq \mathcal{O}^\star$; moreover, for $j \in S(c^\star) \setminus \{w^\star\}$, $w_j(o) = \mathbb{P}(w_\ell = j \mid c^\star, r)/\mathbb{P}(w_\ell = w^\star \mid c^\star, r) \leq \Gamma_2$ by (H2). Since $\mathcal{G}$ contains the true context, $S_{\mathcal{G}} = S(c^\star) \setminus \{w^\star\}$, of size $m = n_{c^\star} - 1 \geq 1$, so $s_{m+1} = s_{n_{c^\star}}$ and $\Gamma_{\mathcal{G}} \leq \Gamma_2$.

By the proof of [Proposition 4.18](#), $\mathbb{P}(o \mid \mathbf{I}_t) > 0$ iff o has positive prior mass, and for $o \in \mathcal{O}^\star$, $\mathbb{P}(w_\ell = w^\star \mid o, \mathbf{I}_t) = \pi_\star(o)$. The contexts of positive posterior mass with $c_\ell = c^\star$ are exactly those in $\mathcal{G}$, and the token is correct iff $c_\ell = c^\star$ and $w_\ell = w^\star$, so

$$\pi_\star^{\text{tok}} = \sum_{o \in \mathcal{G}} \mathbb{P}(o \mid \mathbf{I}_t) \pi_\star(o), \quad \mathbb{P}(o \notin \mathcal{G} \mid \mathbf{I}_t) = \mathbb{P}(c_\ell \neq c^\star \mid \mathbf{I}_t) = 1 - \pi_\star^{\ell,1}.$$

Therefore, as in the proof of [Proposition 4.18](#),

$$1 - \pi_\star^{\text{tok}} = \mathbb{P}(o \notin \mathcal{G} \mid \mathbf{I}_t) + \sum_{o \in \mathcal{G}} \mathbb{P}(o \mid \mathbf{I}_t) (1 - \pi_\star(o)) \leq (1 - \pi_\star^{\ell,1}) + \sup_{o \in \mathcal{G}} (1 - \pi_\star(o)).$$

By [Eq. \(81\)](#) and the proof of [Theorem 4.19\(a\)](#), conditionally on $\mathbf{x}_1$ and with probability at least $1 - 2e^{-u^2/2} - e^{-u'}$, $\sup_{o \in \mathcal{G}} (1 - \pi_\star(o)) \leq \Gamma_{\mathcal{G}} \exp(u' - \rho_2(\rho_2 - s_{n_{c^\star}} - 2u))$, and $\Gamma_{\mathcal{G}} \leq \Gamma_2$. This gives (ii); averaging over $\mathbf{x}_1$ gives the unconditional probability bound.

(iii) Here $\mathcal{O}^\star = \{(c, r) : \mathbb{P}(c_\ell = c, w_\ell = w^\star, r) > 0\}$; every $(c, r) \in \mathcal{O}^\star$ has positive mass and $w^\star \in S(c)$ (validity), so $S((c, r)) = S(c)$ by [Eq. \(82\)](#). Hence, for every $o = (c, r) \in \mathcal{O}^\star$,

$$J' = \bigcap_{c'' : w^\star \in S(c'')} S(c'') \setminus \{w^\star\} \subseteq S(c) \setminus \{w^\star\} = S(o) \setminus \{w^\star\},$$

and for $j \in J'$, since $j, w^\star \in S(c)$, (H2) gives $w_j(o) = \mathbb{P}(w_\ell = j \mid c, r)/\mathbb{P}(w_\ell = w^\star \mid c, r) \geq 1/\Gamma_2$. These are the only two properties of J' and Γ' used in the proof of [Theorem 4.19\(b\)](#). Repeating it with Γ_2 in place

of Γ' : $\pi_*(o) \leq (1 + \Gamma_2^{-1} \sum_{j \in J'} e^{\rho_2(g_j - g_{w^*}) - \rho_2^2})^{-1}$ for all $o \in \mathcal{O}^*$; [Theorem 4.15\(b\)](#) with $J = J'$, weights $1/\Gamma_2$ and $m' = |J'| \geq 2$ bounds the right-hand side by $\Gamma_2 e^{-\rho_2 \eta}$ with conditional probability (given $\mathbf{x}_1$) at least $1 - e^{-u^2/2} - e^{-\Lambda_{m'+1}(h)}$; and [Proposition 4.18](#) gives $\pi_*^{\ell,2} \leq \sup_{o \in \mathcal{O}^*} \pi_*(o) \leq \Gamma_2 e^{-\rho_2 \eta}$.

Simultaneity. Let $\delta \in (0, 1)$ and suppose $\rho_1(\rho_1 - s_{v_1} - 2u) \geq \log \Gamma_1 + \bar{D}_2 + u' + \log(1/\delta)$. The right-hand side is at least $u' > 0$ (as $\Gamma_1 \geq 1$ and $\bar{D}_2 \geq 0$), so $\rho_1 > 0$ and $\rho_1 - s_{v_1} - 2u > 0$, i.e. the hypothesis of (i) holds; rearranging,

$$\Gamma_1 e^{\bar{D}_2} \exp(u' - \rho_1(\rho_1 - s_{v_1} - 2u)) = \exp(\log \Gamma_1 + \bar{D}_2 + u' - \rho_1(\rho_1 - s_{v_1} - 2u)) \leq e^{-\log(1/\delta)} = \delta,$$

so $1 - \pi_*^{\ell,1} \leq \delta$ on the event of (i). The condition on $\rho_2(t)$ is the hypothesis of (iii). By the union bound, both events hold with probability at least $1 - 5e^{-u^2/2} - e^{-u'} - e^{-\Lambda_{m'+1}(h)}$, and on their intersection both statements hold at the same t . $\square$

Proof of [Corollary 4.21](#)

Proof. Idea. By [Proposition 5.5\(ii\)](#) the anchor error is $2\mathbb{E}[1 - \pi_*^{\ell,k}]$, so it suffices to bound the expected posterior error. The high-probability bounds of [Corollary 4.20](#) turn into bounds in expectation because $1 - \pi_*$ lies in $[0, 1]$: on the good event we use the bound, and off it we pay at most the probability of the bad event.

The identity $\mathbb{E}\|\mathbf{p}^{\ell,k}(\mathbf{I}_t) - \mathbf{x}_1^{\ell,k}\|_1 = 2\mathbb{E}[1 - \pi_*^{\ell,k}]$ in [Proposition 5.5\(ii\)](#) holds for any μ_1 (its proof uses only that $\mathbf{p}^{\ell,k}$ is a probability vector), so it remains to bound $\mathbb{E}[1 - \pi_*]$. We use, as in the proof of [Proposition 5.5\(ii\)](#): if $0 \leq X \leq 1$ and $X \leq b$ on an event $\mathcal{E}$, then $X \leq b\mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{E}^c}$, hence $\mathbb{E}X \leq b + \mathbb{P}(\mathcal{E}^c)$ (for $b \geq 0$). All bounds below hold conditionally on $\mathbf{x}_1$ (see the proof of [Corollary 4.20](#)), which fixes the true token (c^*, w^*) and hence n_{c^*} ; the token bound, which involves n_{c^*} , is to be read in this sense, and averaging over $\mathbf{x}_1$ gives the unconditional version with the second term averaged over c^* .

Cluster. If $\rho_1 \geq s_{v_1} + 2u$, apply the fact above with $X = 1 - \pi_*^{\ell,1}$ and the event of [Corollary 4.20\(i\)](#), whose complement has probability at most $4e^{-u^2/2} + e^{-u'}$. If $0 \leq \rho_1 < s_{v_1} + 2u$, then $\rho_1(\rho_1 - s_{v_1} - 2u) \leq 0$, so the claimed bound is at least $\Gamma_1 e^{\bar{D}_2} e^{u'} \geq 1 \geq \mathbb{E}[1 - \pi_*^{\ell,1}]$ and holds trivially.

Token. Suppose $n_{c^*} \geq 2$ and $\rho_2 \geq s_{n_{c^*}} + 2u$, and let $\mathcal{E}$ be the event of [Corollary 4.20\(ii\)](#), with $\mathbb{P}(\mathcal{E}^c | \mathbf{x}_1) \leq 2e^{-u^2/2} + e^{-u'}$. On $\mathcal{E}$, $1 - \pi_*^{\text{tok}} \leq (1 - \pi_*^{\ell,1}) + \Gamma_2 e^{u' - \rho_2(\rho_2 - s_{n_{c^*}} - 2u)}$; on $\mathcal{E}^c$, $1 - \pi_*^{\text{tok}} \leq 1 \leq (1 - \pi_*^{\ell,1}) + 1$ because $1 - \pi_*^{\ell,1} \geq 0$. Hence

$$1 - \pi_*^{\text{tok}} \leq (1 - \pi_*^{\ell,1}) + \Gamma_2 e^{u' - \rho_2(\rho_2 - s_{n_{c^*}} - 2u)} \mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{E}^c},$$

and taking conditional expectations, $\mathbb{E}[1 - \pi_*^{\text{tok}} | \mathbf{x}_1] \leq \mathbb{E}[1 - \pi_*^{\ell,1} | \mathbf{x}_1] + \Gamma_2 e^{u' - \rho_2(\rho_2 - s_{n_{c^*}} - 2u)} + 2e^{-u^2/2} + e^{-u'}$. If $\rho_2 < s_{n_{c^*}} + 2u$, the term $\Gamma_2 e^{u' - \rho_2(\rho_2 - s_{n_{c^*}} - 2u)} \geq e^{u'} > 1$ makes the bound trivial. If $n_{c^*} = 1$, then $S(o) \subseteq S(c^*) = \{w^*\}$ for every $o \in \mathcal{G}$, so $\pi_*(o) = 1$ and the identity in the proof of [Corollary 4.20\(ii\)](#) gives $\pi_*^{\text{tok}} = \sum_{o \in \mathcal{G}} \mathbb{P}(o | \mathbf{I}_t) = \pi_*^{\ell,1}$; the second term is then absent. $\square$

Proofs of [Remarks 4.22](#) and [4.23](#)

Proof. Idea. Both remarks evaluate the constants of the general results for a specific μ_1 . For product data the context factor is common to N_j and N_{i^*} and cancels, so the weights are the prior odds p_j/p_{i^*} , and the shifted threshold is the root of a quadratic. For a Markov chain the conditional law of one token given all others depends only on its two neighbours, so the ratios in (H1) and (H2) are controlled by the range $[a, b]$ of the transition probabilities.

Zipf. For product data, $\pi(i, o) = p_i \pi_{-B}(o)$ with π_{-B} the law of the context. In [Eq. \(23\)](#),

$$\Delta_j = \log \frac{p_j \sum_o \pi_{-B}(o) L(o)}{p_{i^*} \sum_o \pi_{-B}(o) L(o)} = \log \frac{p_j}{p_{i^*}},$$

the common factor being positive ($L > 0$ and π_{-B} is a probability vector). Here $S_B = \{i : p_i > 0\}$, so $\sum_{j \in S_B \setminus \{i^*\}} w_j = \sum_{j \in S_B \setminus \{i^*\}} p_j/p_{i^*} = (1 - p_{i^*})/p_{i^*}$ and, with $m = |S_B| - 1$, $\bar{w} = (1 - p_{i^*})/(m p_{i^*})$. The crossover ρ_w solves $\bar{w} e^{-\rho(\rho - s)} = 1$, i.e. $\rho^2 - s\rho - \log \bar{w} = 0$, whose larger root is $\rho_w = \frac{1}{2}(s + \sqrt{s^2 + 4 \log \bar{w}})$

(when $s^2 + 4\log \bar{w} \geq 0$). With $\lambda := \log \bar{w}$ fixed and $s \rightarrow \infty$, the expansion $\sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8} + O(x^3)$ at $x = 4\lambda/s^2$ gives $\sqrt{s^2 + 4\lambda} = s\sqrt{1 + 4\lambda/s^2} = s + 2\lambda/s - 2\lambda^2/s^3 + O(s^{-5})$, hence $\rho_w = s + \lambda/s + O(s^{-3})$.

Markov. Let the tokens $x_1, \dots, x_L$ form a Markov chain with initial law ν and transition matrix P , all of whose entries lie in $[a, b]$, so $\mathbb{P}(x_1, \dots, x_L) = \nu(x_1) \prod_{k=1}^{L-1} P(x_k, x_{k+1})$. Let $1 < \ell < L$ and let $r = (x_k)_{k \neq \ell}$ have positive mass. Exactly two factors contain x_ℓ ; the others form a constant $C(r) > 0$, so

$$\mathbb{P}(x_\ell = \tau \mid r) = \frac{C(r)P(x_{\ell-1}, \tau)P(\tau, x_{\ell+1})}{\sum_{\tau'} C(r)P(x_{\ell-1}, \tau')P(\tau', x_{\ell+1})} = \frac{q_\tau}{\sum_{\tau'} q_{\tau'}}, \quad q_\tau := P(x_{\ell-1}, \tau)P(\tau, x_{\ell+1}) \in [a^2, b^2].$$

Identify token τ with its code (c, w) , $w \in S(c)$; we assume every cluster is used, $n_{\min} \geq 1$ (otherwise $n_{\max}/n_{\min} = \infty$ and there is nothing to prove). Then $\mathbb{P}(c_\ell = c \mid r) = \sum_{w \in S(c)} q_{(c,w)} / \sum_{\tau'} q_{\tau'}$ with numerator in $[n_c a^2, n_c b^2]$, so for any clusters c, c'

$$\frac{\mathbb{P}(c_\ell = c \mid r)}{\mathbb{P}(c_\ell = c' \mid r)} \leq \frac{n_c b^2}{n_{c'} a^2} \leq \frac{n_{\max}}{n_{\min}} \left(\frac{b}{a}\right)^2,$$

which bounds Γ_1 . Within cluster c , $\mathbb{P}(w_\ell = w \mid c_\ell = c, r) = q_{(c,w)} / \sum_{w' \in S(c)} q_{(c,w')}$, so for $w, w' \in S(c)$ the ratio is $q_{(c,w)} / q_{(c,w')} \leq b^2 / a^2$, which bounds Γ_2 . $\square$

Scope of the bounds. (1) The constants are worst case over contexts. Language has hard constraints ($\pi(i \mid o) = 0$ for most i); [Theorem 4.19](#) handles them by shrinking the competitor set $\mathcal{G}_c$, but (H1)–(H2) and rectangular support fail whenever a symbol of the block is impossible in some context of positive mass, and then only the mixture form with a smaller $\mathcal{G}$ applies, whose context term must be controlled separately. (2) The cluster bound pays e^{D_2} for the evidence of the word block; it is informative only while the word level is still noisy (ρ_2 small), which is the staggered regime, and not if the word level is resolved first. (3) The results are at a fixed interpolation time and per position (with union bounds over positions); they say nothing about the autonomous flow itself.

D Proofs for Section 5

Lemma D.1 (L^2 projection). *Let $\mathbf{z}$ be a random vector, $\mathbf{Y}$ a random vector with $\mathbb{E}\|\mathbf{Y}\|^2 < \infty$, and $\mathbf{m}(\mathbf{z}) := \mathbb{E}[\mathbf{Y} \mid \mathbf{z}]$. For every measurable $\mathbf{f}$ with $\mathbb{E}\|\mathbf{f}(\mathbf{z})\|^2 < \infty$,*

$$\mathbb{E}\|\mathbf{f}(\mathbf{z}) - \mathbf{Y}\|^2 = \mathbb{E}\|\mathbf{f}(\mathbf{z}) - \mathbf{m}(\mathbf{z})\|^2 + \mathbb{E}\|\mathbf{m}(\mathbf{z}) - \mathbf{Y}\|^2.$$

In particular $\mathbf{f} = \mathbf{m}$ almost surely is the unique minimiser.

Proof. Idea. The identity is Pythagoras' theorem in L^2 : the error $\mathbf{m}(\mathbf{z}) - \mathbf{Y}$ of the conditional mean is orthogonal to every square-integrable function of $\mathbf{z}$, so the cross term in the expansion of the square vanishes. The tower property gives this orthogonality; the rest of the argument checks that every term is finite, which is why it begins by showing that $\mathbf{m}(\mathbf{z})$ is square integrable.

By the Doob–Dynkin lemma a version of $\mathbb{E}[\mathbf{Y} \mid \mathbf{z}]$ is a measurable function $\mathbf{m}$ of $\mathbf{z}$. The map $\mathbf{y} \mapsto \|\mathbf{y}\|^2$ is convex, so the conditional Jensen inequality gives $\|\mathbf{m}(\mathbf{z})\|^2 \leq \mathbb{E}[\|\mathbf{Y}\|^2 \mid \mathbf{z}]$ almost surely, and taking expectations, $\mathbb{E}\|\mathbf{m}(\mathbf{z})\|^2 \leq \mathbb{E}\|\mathbf{Y}\|^2 < \infty$. Put $\mathbf{g} := \mathbf{f}(\mathbf{z}) - \mathbf{m}(\mathbf{z})$ and $\mathbf{h} := \mathbf{m}(\mathbf{z}) - \mathbf{Y}$; by $\|\mathbf{a} + \mathbf{a}'\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{a}'\|^2$ both are square integrable. Expanding $\mathbf{f}(\mathbf{z}) - \mathbf{Y} = \mathbf{g} + \mathbf{h}$,

$$\mathbb{E}\|\mathbf{f}(\mathbf{z}) - \mathbf{Y}\|^2 = \mathbb{E}\|\mathbf{g}\|^2 + \mathbb{E}\|\mathbf{h}\|^2 + 2\mathbb{E}\langle \mathbf{g}, \mathbf{h} \rangle = \mathbb{E}\|\mathbf{f}(\mathbf{z}) - \mathbf{m}(\mathbf{z})\|^2 + \mathbb{E}\|\mathbf{m}(\mathbf{z}) - \mathbf{Y}\|^2 + 2 \sum_i \mathbb{E}[g_i h_i],$$

where each product $g_i h_i$ is integrable by the Cauchy–Schwarz inequality, $\mathbb{E}[g_i h_i] \leq (\mathbb{E}g_i^2)^{1/2} (\mathbb{E}h_i^2)^{1/2}$. Each g_i is a function of $\mathbf{z}$, so by the tower property and the pull-out property,

$$\mathbb{E}[g_i h_i] = \mathbb{E}[\mathbb{E}[g_i h_i \mid \mathbf{z}]] = \mathbb{E}[g_i \mathbb{E}[h_i \mid \mathbf{z}]] = \mathbb{E}[g_i (m_i(\mathbf{z}) - \mathbb{E}[Y_i \mid \mathbf{z}])] = 0,$$

since $\mathbb{E}[Y_i | \mathbf{z}] = m_i(\mathbf{z})$. This proves the identity. The second term on its right does not depend on $\mathbf{f}$, and $\mathbf{f} = \mathbf{m}$ is admissible because $\mathbb{E}\|\mathbf{m}(\mathbf{z})\|^2 < \infty$; the first term is nonnegative and vanishes iff $\mathbf{f}(\mathbf{z}) = \mathbf{m}(\mathbf{z})$ almost surely. $\square$

The regression target contains the velocity $\dot{\mathbf{I}}_t$ multiplied by a function of $\mathbf{I}_t$ that need not be bounded or integrable on its own. Its conditional mean is computed through a regular conditional law.

Lemma D.2 (Conditional mean of an affine function of the velocity). *Let $\mathbf{c} : \mathbb{R}^D \rightarrow \mathbb{R}^m$ and $\mathbf{G} : \mathbb{R}^D \rightarrow \mathbb{R}^{m \times D}$ be Borel and assume $\mathbb{E}\|\mathbf{c}(\mathbf{I}_t) + \mathbf{G}(\mathbf{I}_t)\dot{\mathbf{I}}_t\| < \infty$, under the joint law of $t \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0$, $\mathbf{x}_1$. Then*

$$\mathbb{E}[\mathbf{c}(\mathbf{I}_t) + \mathbf{G}(\mathbf{I}_t)\dot{\mathbf{I}}_t | \mathbf{I}_t] = \mathbf{c}(\mathbf{I}_t) + \mathbf{G}(\mathbf{I}_t)\mathbf{b}(\mathbf{I}_t) \quad \text{almost surely.} \quad (83)$$

Proof. Idea. If $\mathbf{G}(\mathbf{I}_t)\dot{\mathbf{I}}_t$ were integrable on its own, the claim would follow by pulling the $\mathbf{I}_t$ -measurable factor $\mathbf{G}(\mathbf{I}_t)$ out of the conditional expectation. Only the sum $\mathbf{c}(\mathbf{I}_t) + \mathbf{G}(\mathbf{I}_t)\dot{\mathbf{I}}_t$ is assumed integrable, so we pass to a regular conditional law of $\dot{\mathbf{I}}_t$ given $\mathbf{I}_t$: for a fixed value $\mathbf{x}$ the conditional expectation becomes an integral against a probability measure, in which $\mathbf{G}(\mathbf{x})$ is a constant matrix and linearity of the integral gives the result.

$(\mathbf{I}_t, \dot{\mathbf{I}}_t)$ is a random vector in $\mathbb{R}^D \times \mathbb{R}^D$, so by the disintegration theorem there is a probability kernel κ from $\mathbb{R}^D$ to $\mathbb{R}^D$ such that $\kappa(\mathbf{I}_t, \cdot)$ is a regular conditional law of $\dot{\mathbf{I}}_t$ given $\mathbf{I}_t$: for every Borel $\mathbf{h} : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}^m$ with $\mathbb{E}\|\mathbf{h}(\mathbf{I}_t, \dot{\mathbf{I}}_t)\| < \infty$, the integral $\int \mathbf{h}(\mathbf{x}, \mathbf{w}) \kappa(\mathbf{x}, d\mathbf{w})$ converges absolutely for q -a.e. $\mathbf{x}$ and $\mathbb{E}[\mathbf{h}(\mathbf{I}_t, \dot{\mathbf{I}}_t) | \mathbf{I}_t] = \int \mathbf{h}(\mathbf{I}_t, \mathbf{w}) \kappa(\mathbf{I}_t, d\mathbf{w})$ almost surely. With $\mathbf{h}(\mathbf{x}, \mathbf{w}) = \mathbf{w}$, which is integrable by Lemma B.3(i), and Eq. (1), there is a set $N \subset \mathbb{R}^D$ with $\mathbb{P}(\mathbf{I}_t \in N) = 0$ such that for $\mathbf{x} \notin N$

$$\int \|\mathbf{w}\| \kappa(\mathbf{x}, d\mathbf{w}) < \infty \quad \text{and} \quad \int \mathbf{w} \kappa(\mathbf{x}, d\mathbf{w}) = \mathbf{b}(\mathbf{x})$$

(two versions of $\mathbb{E}[\dot{\mathbf{I}}_t | \mathbf{I}_t]$ agree almost surely). Now take $\mathbf{h}(\mathbf{x}, \mathbf{w}) = \mathbf{c}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{w}$, which is integrable by hypothesis. For $\mathbf{x} \notin N$, $\|\mathbf{c}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{w}\| \leq \|\mathbf{c}(\mathbf{x})\| + \|\mathbf{G}(\mathbf{x})\| \|\mathbf{w}\|$ is $\kappa(\mathbf{x}, \cdot)$ -integrable, and since $\kappa(\mathbf{x}, \cdot)$ is a probability measure and $\mathbf{w} \mapsto \mathbf{G}(\mathbf{x})\mathbf{w}$ is linear,

$$\int (\mathbf{c}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{w}) \kappa(\mathbf{x}, d\mathbf{w}) = \mathbf{c}(\mathbf{x}) + \mathbf{G}(\mathbf{x}) \int \mathbf{w} \kappa(\mathbf{x}, d\mathbf{w}) = \mathbf{c}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{b}(\mathbf{x}).$$

Evaluating at $\mathbf{x} = \mathbf{I}_t \notin N$ gives Eq. (83). $\square$

Since $q > 0$ on $\mathbb{R}^D$ (Lemma B.3(ii)) and q is the density of $\mathbf{I}_t$, a Borel set N satisfies $\mathbb{P}(\mathbf{I}_t \in N) = \int_N q d\mathbf{x} = 0$ iff N is Lebesgue-null. Hence

$$\text{a statement about } \mathbf{I}_t \text{ holds almost surely} \iff \text{it holds for Lebesgue-almost every } \mathbf{x}. \quad (84)$$

D.1 Proof of Proposition 5.1

Proof. Idea. The objective is a least-squares regression of $\mathbf{Y} = \bar{\mathbf{T}}(\mathbf{I}_t) + \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{I}_t)\dot{\mathbf{I}}_t$ on $\mathbf{I}_t$, so its minimiser is $\mathbb{E}[\mathbf{Y} | \mathbf{I}_t]$ (Lemma D.1). The target is affine in the velocity, and the conditional mean of the velocity given $\mathbf{I}_t$ is $\mathbf{b}$ by definition, so Lemma D.2 gives Eq. (26). The blockwise form holds because the squared norm is a sum over blocks, and the fixed-point statement reduces to the fact that a function continuous off the finite set $\mathcal{M}_1$ that vanishes almost everywhere vanishes everywhere off $\mathcal{M}_1$.

Minimiser. Put $\mathbf{z} := \mathbf{I}_t$ (under the joint law of $t \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0$, $\mathbf{x}_1$) and $\mathbf{Y} := \bar{\mathbf{T}}(\mathbf{I}_t) + \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{I}_t)\dot{\mathbf{I}}_t$, which is square integrable by assumption. The objective is $\mathbb{E}\|\mathbf{T}(\mathbf{z}) - \mathbf{Y}\|^2$, so by Lemma D.1 its unique minimiser over measurable $\mathbf{T}$ with $\mathbb{E}\|\mathbf{T}(\mathbf{I}_t)\|^2 < \infty$ is $\mathbf{T}^* = \mathbf{m}$, $\mathbf{m}(\mathbf{I}_t) = \mathbb{E}[\mathbf{Y} | \mathbf{I}_t]$, up to $\mathbb{P}$ -null sets of values of $\mathbf{I}_t$. Lemma D.2 with $\mathbf{c} = \bar{\mathbf{T}}$ and $\mathbf{G} = \mathbf{J}_{\bar{\mathbf{T}}}$ (both continuous, hence Borel, as $\bar{\mathbf{T}} \in C^1$; integrability holds because $\mathbb{E}\|\mathbf{Y}\| \leq (\mathbb{E}\|\mathbf{Y}\|^2)^{1/2} < \infty$) gives

$$\mathbf{m}(\mathbf{I}_t) = \mathbb{E}[\bar{\mathbf{T}}(\mathbf{I}_t) + \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{I}_t)\dot{\mathbf{I}}_t | \mathbf{I}_t] = \bar{\mathbf{T}}(\mathbf{I}_t) + \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{I}_t)\mathbf{b}(\mathbf{I}_t) \quad \text{almost surely,}$$

and by Eq. (84) this is Eq. (26) for Lebesgue-almost every $\mathbf{x}$. The conditioning is on $\mathbf{I}_t$ under the law in which t is uniform, which is exactly the law used to define $\mathbf{b}$ in Eq. (1).

Blocks. $\|\mathbf{v}\|^2 = \sum_{\ell,k} \|\mathbf{v}^{\ell,k}\|^2$ for every $\mathbf{v} \in \mathbb{R}^D$, so

$$\mathbb{E}\|\mathbf{T}(\mathbf{I}_t) - \mathbf{Y}\|^2 = \sum_{\ell,k} \mathbb{E}\|\mathbf{T}^{\ell,k}(\mathbf{I}_t) - \mathbf{Y}^{\ell,k}\|^2, \quad \mathbf{Y}^{\ell,k} = \bar{\mathbf{T}}^{\ell,k}(\mathbf{I}_t) + \mathbf{J}_{\bar{\mathbf{T}}^{\ell,k}}(\mathbf{I}_t)\dot{\mathbf{I}}_t,$$

where $\mathbf{J}_{\bar{\mathbf{T}}^{\ell,k}}$ is the $v_k \times D$ block of rows of $\mathbf{J}_{\bar{\mathbf{T}}}$, and $\mathbb{E}\|\mathbf{T}(\mathbf{I}_t)\|^2 < \infty$ iff $\mathbb{E}\|\mathbf{T}^{\ell,k}(\mathbf{I}_t)\|^2 < \infty$ for every block. The (ℓ, k) term depends only on $\mathbf{T}^{\ell,k}$, so minimising the sum over $\mathbf{T} = (\mathbf{T}^{\ell,k})$ is the same as minimising each term separately. Lemmas D.1 and D.2 applied to block (ℓ, k) give $\mathbf{T}^{*,k} = \bar{\mathbf{T}}^{\ell,k} + \mathbf{J}_{\bar{\mathbf{T}}^{\ell,k}}\mathbf{b}$ almost everywhere, which is block (ℓ, k) of Eq. (26). With $\bar{\mathbf{T}} = \mathbf{T}_\theta$ frozen, the summand of Eq. (25) has the value $\|\mathbf{T}^{\ell,k}(\mathbf{I}_t) - \mathbf{Y}^{\ell,k}\|^2$ at $\mathbf{T} = \mathbf{T}_\theta$, so Eq. (25) has the same minimiser as the unblocked DBTM loss.

Fixed points. $\mathbf{T}^*$ is determined up to Lebesgue-null sets, so $\mathbf{T}^* = \bar{\mathbf{T}}$ means equality almost everywhere, and by Eq. (26) this holds iff $\mathbf{J}_{\bar{\mathbf{T}}}\mathbf{b} = \mathbf{0}$ almost everywhere. The set $\mathcal{M}_1$ is finite, so $U := \mathbb{R}^D \setminus \mathcal{M}_1$ is open and $\mathcal{M}_1$ is Lebesgue-null. On U , $\mathbf{J}_{\bar{\mathbf{T}}}$ is continuous ($\bar{\mathbf{T}} \in C^1$) and $\mathbf{b}$ is continuous (Proposition 3.3), so $\mathbf{g} := \mathbf{J}_{\bar{\mathbf{T}}}\mathbf{b}$ is continuous on U . If $\mathbf{g} = \mathbf{0}$ almost everywhere and $\mathbf{g}(\mathbf{x}_o) \neq \mathbf{0}$ for some $\mathbf{x}_o \in U$, then by continuity $\mathbf{g} \neq \mathbf{0}$ on a ball $B(\mathbf{x}_o, r) \subset U$, which has positive measure, a contradiction; so $\mathbf{g} = \mathbf{0}$ on U . Conversely, $\mathbf{g} = \mathbf{0}$ on U implies $\mathbf{g} = \mathbf{0}$ almost everywhere because $\mathcal{M}_1$ is null. Hence the fixed points are exactly the C^1 maps with $\mathbf{J}_{\bar{\mathbf{T}}}\mathbf{b} = \mathbf{0}$ on $\mathbb{R}^D \setminus \mathcal{M}_1$, which is the first line of Eq. (2). $\square$

D.2 Proof of Proposition 5.5

Proof. Idea. For (i), the expected cross-entropy given $\mathbf{I}_t = \mathbf{x}$ is the entropy of the posterior, which does not depend on $\mathbf{q}$, plus the Kullback–Leibler divergence from the posterior to $\mathbf{q}(\mathbf{x})$, which vanishes only at $\mathbf{q}(\mathbf{x}) = \mathbf{p}^{\ell,k}(\mathbf{x})$; this is the usual argument that cross-entropy is minimised by the true conditional law. For (ii), the ℓ_1 distance from a probability vector to the vertex of the true symbol is $2(1 - \pi_\star^{\ell,k})$, so the gap depends only on the law of $\pi_\star^{\ell,k}$, which Theorem 4.2 controls on both sides of the threshold. Its high-probability statements become bounds on the mean because $0 \leq 1 - \pi_\star^{\ell,k} \leq 1$.

Throughout t is fixed. Conditionally on t , $\mathbf{I}_t$ has the density $\sum_j \pi_j p(\cdot | t, \mathbf{j})$, which is strictly positive (Lemma B.1), so the analogue of Eq. (84) holds: almost surely in $\mathbf{I}_t$ is the same as Lebesgue-almost everywhere. Write J for the symbol of block (ℓ, k) , $\mathbf{x}_1^{\ell,k} = \mathbf{e}_J$, so that $p_i^{\ell,k}(\mathbf{x}) = \mathbb{P}(J = i | \mathbf{I}_t = \mathbf{x})$.

(i) Let $\mathbf{q}$ be a measurable map into the open simplex of $\mathbb{R}^{v_k}$. Then $\text{CE}(\mathbf{q}(\mathbf{x}), \mathbf{e}_i) = -\log q_i(\mathbf{x}) \in [0, \infty)$, so $\text{CE}(\mathbf{q}(\mathbf{I}_t), \mathbf{x}_1^{\ell,k}) = \sum_i \mathbb{1}_{[J=i]} (-\log q_i(\mathbf{I}_t))$ is nonnegative, and conditioning on $\mathbf{I}_t$ (the factors $-\log q_i(\mathbf{I}_t)$ are functions of $\mathbf{I}_t$; for nonnegative variables no integrability is needed),

$$\mathbb{E}[\text{CE}(\mathbf{q}(\mathbf{I}_t), \mathbf{x}_1^{\ell,k}) | \mathbf{I}_t = \mathbf{x}] = -\sum_i p_i \log q_i = -\sum_i p_i \log p_i + \sum_i p_i \log \frac{p_i}{q_i} = H(\mathbf{p}) + \text{KL}(\mathbf{p} \| \mathbf{q}), \quad (85)$$

with $\mathbf{p} = \mathbf{p}^{\ell,k}(\mathbf{x})$, $\mathbf{q} = \mathbf{q}(\mathbf{x})$ and the convention $0 \log 0 = 0$; every term is finite because $q_i > 0$. By Gibbs' inequality $\text{KL}(\mathbf{p} \| \mathbf{q}) \geq 0$, with equality iff $\mathbf{q} = \mathbf{p}$: with $S = \{i : p_i > 0\}$ and the strict concavity of $\log$ (Jensen),

$$-\text{KL}(\mathbf{p} \| \mathbf{q}) = \sum_{i \in S} p_i \log \frac{q_i}{p_i} \leq \log \sum_{i \in S} q_i \leq \log 1 = 0,$$

with equality in the first step iff q_i/p_i is constant on S and in the second iff $\sum_{i \in S} q_i = 1$; together these force $q_i = p_i$ for all i . Taking expectations in Eq. (85) (tower property for nonnegative variables),

$$\mathbb{E} \text{CE}(\mathbf{q}(\mathbf{I}_t), \mathbf{x}_1^{\ell,k}) = \mathbb{E} H(\mathbf{p}^{\ell,k}(\mathbf{I}_t)) + \mathbb{E} \text{KL}(\mathbf{p}^{\ell,k}(\mathbf{I}_t) \| \mathbf{q}(\mathbf{I}_t)), \quad 0 \leq \mathbb{E} H(\mathbf{p}^{\ell,k}(\mathbf{I}_t)) \leq \log v_k.$$

The first term is finite and does not depend on $\mathbf{q}$, and the second is nonnegative. Hence $\mathbb{E} \text{CE}$ is minimal iff $\mathbb{E} \text{KL}(\mathbf{p}^{\ell,k}(\mathbf{I}_t) \| \mathbf{q}(\mathbf{I}_t)) = 0$, i.e. iff $\text{KL}(\mathbf{p}^{\ell,k}(\mathbf{I}_t) \| \mathbf{q}(\mathbf{I}_t)) = 0$ almost surely, i.e. iff $\mathbf{q} = \mathbf{p}^{\ell,k}$ almost everywhere. This identifies the minimiser when $\mathbf{p}^{\ell,k}$ takes values in the open simplex, which holds iff every symbol of block (ℓ, k) has positive marginal probability, since $p_i^{\ell,k}(\mathbf{x}) = \sum_{j: j_{\ell,k}=i} \pi_j p(\mathbf{x} | t, \mathbf{j}) / \sum_j \pi_j p(\mathbf{x} | t, \mathbf{j})$ and every Gaussian density is positive; in particular under the assumptions of Theorem 4.5.

(ii) Let $i^* = J$ be the true symbol, $\mathbf{p} = \mathbf{p}^{\ell,k}(\mathbf{I}_t)$ and $\pi_{\star}^{\ell,k} := p_{i^*}$. Since $\sum_i p_i = 1$ and $p_i \geq 0$,

$$\|\mathbf{p} - \mathbf{e}_{i^*}\|_1 = (1 - p_{i^*}) + \sum_{i \neq i^*} p_i = (1 - p_{i^*}) + (1 - p_{i^*}) = 2(1 - \pi_{\star}^{\ell,k}),$$

and taking expectations gives the identity. Under the assumptions of [Theorem 4.5](#), $\pi_{\star}^{\ell,k}$ has the law of [Eq. \(11\)](#) with $n = v_k$ and $\rho = \rho_k := \rho_k(t)$, a deterministic number because t is fixed, so [Theorem 4.2](#) applies to $\pi_{\star}^{\ell,k}$.

Upper bound. Let $\rho_k \geq s_{v_k} + 2u$, $\varepsilon_a := e^{u' - \rho_k(\rho_k - s_{v_k} - 2u)}$ and $\mathcal{G} := \{1 - \pi_{\star}^{\ell,k} \leq \varepsilon_a\}$. By [Theorem 4.2\(a\)](#), $\mathbb{P}(\mathcal{G}^c) \leq 2e^{-u^2/2} + e^{-u'}$. Since $0 \leq 1 - \pi_{\star}^{\ell,k} \leq 1$,

$$\mathbb{E}[1 - \pi_{\star}^{\ell,k}] = \mathbb{E}[(1 - \pi_{\star}^{\ell,k})\mathbb{1}_{\mathcal{G}}] + \mathbb{E}[(1 - \pi_{\star}^{\ell,k})\mathbb{1}_{\mathcal{G}^c}] \leq \varepsilon_a \mathbb{P}(\mathcal{G}) + \mathbb{P}(\mathcal{G}^c) \leq e^{u' - \rho_k(\rho_k - s_{v_k} - 2u)} + 2e^{-u^2/2} + e^{-u'}.$$

Lower bound. Let $\rho_k \leq s_{v_k} - h - u - \eta$ and $\mathcal{H} := \{\pi_{\star}^{\ell,k} \leq e^{-\rho_k \eta}\}$. By [Theorem 4.2\(b\)](#), $\mathbb{P}(\mathcal{H}) \geq 1 - e^{-u^2/2} - e^{-\Lambda_{v_k}(h)}$. On $\mathcal{H}$, $1 - \pi_{\star}^{\ell,k} \geq 1 - e^{-\rho_k \eta} \geq 0$ (as $\rho_k \geq 0$, $\eta > 0$), and $1 - \pi_{\star}^{\ell,k} \geq 0$ everywhere, so

$$\mathbb{E}[1 - \pi_{\star}^{\ell,k}] \geq \mathbb{E}[(1 - \pi_{\star}^{\ell,k})\mathbb{1}_{\mathcal{H}}] \geq (1 - e^{-\rho_k \eta}) \mathbb{P}(\mathcal{H}) \geq (1 - e^{-\rho_k \eta})(1 - e^{-u^2/2} - e^{-\Lambda_{v_k}(h)}),$$

the last step multiplying the lower bound on $\mathbb{P}(\mathcal{H})$ by the nonnegative factor $1 - e^{-\rho_k \eta}$. $\square$

D.3 Proof of [Proposition 5.6](#)

Proof. Idea. Since t is uniform and independent of $(\mathbf{x}_0, \mathbf{x}_1)$, the probability of an event is the integral over t of its probability at fixed t , and at fixed t [Theorem 4.2](#) applies with $\rho = \rho_k(t)$. As ρ_k is increasing, the times at which part (b) applies form an interval $[0, t_k(\cdot)]$ and those at which part (a) applies an interval $[t_k(\cdot), 1]$, so each bound is the length of a time interval times a probability bound that does not depend on t . The limit reuses the parameter choices of [Corollary 4.3](#), and the monotonicity follows by differentiating the explicit formula for t_k^* in the exponent.

Write $s := s_{v_k}$, $\rho_k(t) = ((1 - t)^{-a_k} - 1)/\sigma_k$ and $E_- := \{\pi_{\star}^{\ell,k} \leq e^{-\rho_k(t)\eta}\}$, $E_+ := \{1 - \pi_{\star}^{\ell,k} \leq e^{u' - \rho_k(t)\delta}\}$. Admissibility in [Theorem 4.2\(b\)](#) means $u, \eta > 0$ and $h \in (0, s)$; by hypothesis also $s - h - u - \eta \geq 0$, so that the range $0 \leq \rho \leq s - h - u - \eta$ is nonempty. By [Corollary 4.4](#) with $(a, \sigma) = (a_k, \sigma_k)$, ρ_k is a strictly increasing bijection of $[0, 1]$ onto $[0, \infty)$ with inverse $t_k(\cdot)$, so for $\rho_0 \geq 0$

$$\{t \in [0, 1] : \rho_k(t) \leq \rho_0\} = [0, t_k(\rho_0)], \quad \{t \in [0, 1] : \rho_k(t) \geq \rho_0\} = [t_k(\rho_0), 1], \quad (86)$$

and $t_k(\rho_0) \in [0, 1]$. The variable t is uniform and independent of $(\mathbf{x}_0, \mathbf{x}_1)$, so conditionally on t the pair $(\mathbf{x}_0, \mathbf{x}_1)$ keeps its law, and for every event E

$$\mathbb{P}(E) = \int_0^1 \mathbb{P}(E | t) dt, \quad (87)$$

where $\mathbb{P}(E | t)$ is the probability of E for the interpolant at the fixed time t . By [Theorem 4.5](#), at fixed $t \in (0, 1)$ the block posterior $\pi_{\star}^{\ell,k}$ has the law of [Eq. \(11\)](#) with $n = v_k$ and $\rho = \rho_k(t)$ (the endpoints $t \in \{0, 1\}$ are a null set).

Undecided. Let $t_- := t_k(s - h - u - \eta) \in [0, 1]$. For $t \in (0, t_-]$, [Eq. \(86\)](#) gives $\rho_k(t) \leq s - h - u - \eta$, so [Theorem 4.2\(b\)](#) yields $\mathbb{P}(E_- | t) \geq F_- := 1 - e^{-u^2/2} - e^{-\Lambda_{v_k}(h)}$, a bound that does not depend on t . By [Eq. \(87\)](#) and $\mathbb{P}(E_- | t) \geq 0$ for $t > t_-$,

$$\mathbb{P}(E_-) \geq \int_0^{t_-} \mathbb{P}(E_- | t) dt \geq t_- F_- = t_k(s - h - u - \eta)(1 - e^{-u^2/2} - e^{-\Lambda_{v_k}(h)}).$$

Decided. Let $t_+ := t_k(s + 2u + \delta) \in (0, 1)$. For $t \in [t_+, 1]$, [Eq. \(86\)](#) gives $\rho_k(t) \geq s + 2u + \delta \geq s + 2u$, so [Theorem 4.2\(a\)](#) applies, and since $\rho_k(t) - s - 2u \geq \delta$ and $\rho_k(t) \geq 0$,

$$u' - \rho_k(t)(\rho_k(t) - s - 2u) \leq u' - \rho_k(t)\delta.$$

Hence the event of Eq. (12) is contained in E_+ , and $\mathbb{P}(E_+ | t) \geq F_+ := 1 - 2e^{-u^2/2} - e^{-u'}$. By Eq. (87),

$$\mathbb{P}(E_+) \geq \int_{t_+}^1 \mathbb{P}(E_+ | t) dt \geq (1 - t_+) F_+ = (1 - t_k(s + 2u + \delta))(1 - 2e^{-u^2/2} - e^{-u'}).$$

Limit. Fix $\varepsilon \in (0, 1)$ and take, as in the proof of Corollary 4.3, $h = u = \varepsilon s/4$, $\eta = \varepsilon s/2$ in the first bound and $u = \varepsilon s/4$, $\delta = \varepsilon s/2$, $u' = \log s$ in the second. Then

$$s - h - u - \eta = (1 - \varepsilon)s, \quad s + 2u + \delta = (1 + \varepsilon)s.$$

As $v_k \rightarrow \infty$, $s \rightarrow \infty$, and the proof of Corollary 4.3 shows $e^{-u^2/2} = e^{-\varepsilon^2 s^2/32} \rightarrow 0$, $e^{-\Lambda_{v_k}(h)} \rightarrow 0$ and $e^{-u'} = 1/s \rightarrow 0$; so $F_\pm \rightarrow 1$, while the event thresholds satisfy $e^{-\rho_k \eta} \leq 1$ and $e^{u' - \rho_k \delta} \leq s e^{-(1+\varepsilon)\varepsilon s^2/2} \rightarrow 0$ on the respective time ranges. For the times, with $X_\pm := 1 + \sigma_k(1 \pm \varepsilon)s$ and $X := 1 + \sigma_k s$,

$$|t_k((1 \pm \varepsilon)s) - t_k^*| = |X^{-1/a_k} - X_\pm^{-1/a_k}| \leq X^{-1/a_k} \rightarrow 0, \quad \frac{1 - t_k((1 \pm \varepsilon)s)}{1 - t_k^*} = \left(\frac{X}{X_\pm}\right)^{1/a_k} \rightarrow (1 \pm \varepsilon)^{-1/a_k},$$

since $X_\pm \geq X_- \rightarrow \infty$ and $X/X_\pm \rightarrow (1 \pm \varepsilon)^{-1}$. So both time boundaries tend to t_k^* , and the remaining times $1 - t_k(\cdot)$ agree with $1 - t_k^*$ up to factors that tend to 1 as $\varepsilon \rightarrow 0$.

Monotonicity. $t_k^* = 1 - X^{-1/a_k}$ with $X = 1 + \sigma_k s_{v_k} \geq 1$, and $\partial_a(1 - X^{-1/a}) = -X^{-1/a} \ln X/a^2 \leq 0$, with strict inequality iff $X > 1$, i.e. $v_k \geq 3$. Hence $t_1^* = 1 - (1 + \sigma_1 s_{v_1})^{-1/a_1}$ is decreasing in a_1 , so $a_1 > 1$ gives a smaller coarse budget than $a_1 = 1$; and decreasing a_K below 1 increases t_K^* , while t_1^* depends only on (a_1, σ_1, v_1) and is unchanged. These are the monotonicity statements of Corollary 4.4. $\square$

D.4 Proof of Proposition 5.7

Proof. Idea. All claims are invariants of the bookkeeping in Algorithm 1. For the hierarchy we show that the set of committed and selected blocks stays closed under passing to coarser blocks of the same position: it is closed at the start, and each single addition preserves closedness, so it suffices to check the additions one at a time. Progress and termination rest on one observation, Eq. (88): the frontier is empty only when every remaining block has been selected, so the floor loop can always add a block until $|\Delta| \geq n_r$, and in the last round n_k is the number of remaining blocks.

In this proof k is the number of rounds of Algorithm 1 and levels are indexed by k' . Call a set S of blocks *closed* if $(\ell, k') \in S$ and $k'' < k'$ imply $(\ell, k'') \in S$. Call a block (ℓ, k') *admissible* at a given moment of round r if $(\ell, k') \notin \mathcal{C}_{r-1} \cup \Delta$ and all blocks (ℓ, k'') with $k'' < k'$ lie in $\mathcal{C}_{r-1} \cup \Delta$. The frontier of line 9 is the set of admissible blocks: for a position ℓ whose blocks are not all in $\mathcal{C}_{r-1} \cup \Delta$, the coarsest block not in $\mathcal{C}_{r-1} \cup \Delta$ has all coarser blocks in $\mathcal{C}_{r-1} \cup \Delta$ by minimality, and conversely an admissible block is the coarsest such block of its position. In particular,

$$\text{the frontier is nonempty} \iff \mathcal{C}_{r-1} \cup \Delta \neq \text{all blocks} \iff \Delta \neq \mathcal{R}_r, \quad (88)$$

the last equivalence because $\Delta \subseteq \mathcal{R}_r$ and $\mathcal{R}_r$ is the complement of $\mathcal{C}_{r-1}$.

Hierarchy. If S is closed and every block coarser than $b = (\ell, k')$ lies in S , then $S \cup \{b\}$ is closed. Every block added to Δ is admissible when it is added: the threshold rule (line 6) checks this condition explicitly, and within the loop over k' a block of level k' only requires blocks of levels $< k'$, which have already been processed, so the order of additions within a level is irrelevant; the floor rule (line 9) adds a frontier block. Since $\mathcal{C}_0 = \emptyset$ is closed, induction over the additions shows that $\mathcal{C}_{r-1} \cup \Delta$ is closed at every moment, and $\mathcal{C}_r = \mathcal{C}_{r-1} \cup \Delta$ is closed for every r . Hence whenever a block is committed, every coarser block of its position is committed in the same round or earlier, and was added before it. Committed blocks are never uncommitted, since $\mathcal{C}_r \supseteq \mathcal{C}_{r-1}$.

Progress. For $r \leq k$, $k - r + 1 \geq 1$, so $n_r = \lceil |\mathcal{R}_r|/(k - r + 1) \rceil \leq \lceil |\mathcal{R}_r| \rceil = |\mathcal{R}_r|$. Consider an iteration of the floor loop: it runs only if $|\Delta| < n_r \leq |\mathcal{R}_r|$, so $\Delta \neq \mathcal{R}_r$, the frontier is nonempty by Eq. (88), and one new block is added, increasing $|\Delta|$ by one. Hence the loop stops after at most n_r iterations, and only when $|\Delta| \geq n_r$. In every round $|\Delta| \geq n_r$, which proves the first claim (the alternative “every block is committed” is then not needed).

Termination. In round $r = k$, $n_k = \lceil |\mathcal{R}_k|/1 \rceil = |\mathcal{R}_k|$, so $|\Delta| \geq |\mathcal{R}_k|$ with $\Delta \subseteq \mathcal{R}_k$, i.e. $\Delta = \mathcal{R}_k$, and $\mathcal{C}_k = \mathcal{C}_{k-1} \cup \mathcal{R}_k$ is the set of all blocks. (If $\mathcal{R}_r = \emptyset$ for some earlier r , then $n_r = 0$, $\Delta = \emptyset$, and the remaining rounds commit nothing.)

Evaluations. $\mathbf{T}_\theta$ is evaluated only in line 3, once per round, hence k times. Lines 4–11 are selections, comparisons and Gaussian draws, and the returned decoding (line 13) uses $\hat{\mathbf{x}}$ from the last round without a new evaluation. So the sampler uses exactly k evaluations. $\square$

E Proofs for the Learning Guarantee

The proof of [Theorem 5.2](#) follows the learned map along trajectories of the exact flow, so its bound comes out as an integral against the occupation measure ν of [Eq. \(27\)](#), whereas the transport loss is an expectation under the density q of the interpolant. [Lemma E.1](#) explains why the two measures are comparable: both solve the same stationary continuity equation. The theorem itself uses only the one-sided comparison [Eq. \(93\)](#), which comes from [Theorem 3.8\(iv\)](#) and lets the residual be measured under q .

For $\mathbf{x} \in \mathcal{A}$ ([Theorem 3.8\(ii\)](#)) the trajectory $s \mapsto \mathbf{X}_s(\mathbf{x})$ is C^1 on $[0, \tau(\mathbf{x}))$ with values in Ω , since it solves $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$ with $\mathbf{b}$ continuous on Ω . The field is locally Lipschitz on the open set Ω ($\mathbf{b} \in C^\infty(\Omega)$), so the flow $(\mathbf{x}, s) \mapsto \mathbf{X}_s(\mathbf{x})$ is continuous on its open domain $\{(\mathbf{x}, s) : \mathbf{x} \in \Omega, 0 \leq s < \tau(\mathbf{x})\}$; in particular all integrands $(\mathbf{x}, s) \mapsto f(\mathbf{X}_s(\mathbf{x}))\mathbb{1}_{[s < \tau(\mathbf{x})]}$ with f Borel are jointly Borel, and Tonelli's and Fubini's theorems apply. For Borel $f \geq 0$ on Ω , [Eq. \(27\)](#) and Tonelli give $\int f d\nu = \mathbb{E}_{\mu_0} \int_0^\tau f(\mathbf{X}_s) ds$; if $\int |f| d\nu < \infty$, applying this to f^+ and f^- and subtracting gives

$$\int f d\nu = \mathbb{E}_{\mathbf{x} \sim \mu_0} \int_0^{\tau(\mathbf{x})} f(\mathbf{X}_s(\mathbf{x})) ds, \quad \text{with} \quad \int_0^{\tau(\mathbf{x})} |f(\mathbf{X}_s(\mathbf{x}))| ds < \infty \quad \text{for } \mu_0\text{-a.e. } \mathbf{x}. \quad (89)$$

Lemma E.1 (Both measures solve the same stationary equation). *If $\int \|\mathbf{b}\| d\nu = \mathbb{E}_{\mu_0} \int_0^\tau \|\mathbf{b}(\mathbf{X}_s)\| ds < \infty$, then for every $\phi \in C_c^1(\mathbb{R}^D)$*

$$\int \nabla \phi \cdot \mathbf{b} d\nu = \int \phi d\mu_1 - \int \phi d\mu_0 = \int \nabla \phi(\mathbf{x}) \cdot \mathbf{b}(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}. \quad (90)$$

Proof. Idea. Both equalities express $\int \phi d\mu_1 - \int \phi d\mu_0$ as the expected total change of ϕ along curves that run from a prior sample to a data point. For ν the curves are the trajectories of the flow, along which ϕ changes at the rate $\nabla \phi \cdot \mathbf{b}$; for q they are the interpolant paths, along which ϕ changes at the rate $\nabla \phi \cdot \dot{\mathbf{I}}_t$, and $\dot{\mathbf{I}}_t$ may be replaced by its conditional mean $\mathbf{b}$ because $\nabla \phi(\mathbf{I}_t)$ is a function of $\mathbf{I}_t$. In both cases the fundamental theorem of calculus does the work, and the only delicate point is the limit at the end of the curve, where the trajectory reaches the data; dominated convergence handles it.

Left equality. Let $f := \nabla \phi \cdot \mathbf{b}$ on Ω ; $|f| \leq \|\nabla \phi\|_\infty \|\mathbf{b}\|$, so $\int |f| d\nu < \infty$ by hypothesis and [Eq. \(89\)](#) applies. Fix $\mathbf{x} \in \mathcal{A}$ with $\int_0^{\tau(\mathbf{x})} |f(\mathbf{X}_s)| ds < \infty$; by [Theorem 3.8\(i\)–\(ii\)](#) and [Eq. \(89\)](#) this holds for μ_0 -a.e. $\mathbf{x}$. The map $s \mapsto \phi(\mathbf{X}_s)$ is C^1 on $[0, \tau(\mathbf{x}))$, and by the chain rule

$$\frac{d}{ds} \phi(\mathbf{X}_s) = \nabla \phi(\mathbf{X}_s) \cdot \dot{\mathbf{X}}_s = \nabla \phi(\mathbf{X}_s) \cdot \mathbf{b}(\mathbf{X}_s) = f(\mathbf{X}_s).$$

By the fundamental theorem of calculus, for $s' < \tau(\mathbf{x})$,

$$\phi(\mathbf{X}_{s'}) - \phi(\mathbf{x}) = \int_0^{s'} f(\mathbf{X}_s) ds. \quad (91)$$

Let $s' \uparrow \tau(\mathbf{x})$. On the left, $\mathbf{X}_{s'} \rightarrow \mathbf{T}(\mathbf{x})$ and ϕ is continuous, so the left side tends to $\phi(\mathbf{T}(\mathbf{x})) - \phi(\mathbf{x})$. On the right, $f(\mathbf{X}_s)\mathbb{1}_{[0, s']}(s) \rightarrow f(\mathbf{X}_s)\mathbb{1}_{[0, \tau(\mathbf{x}))}(s)$ pointwise with the integrable dominant $|f(\mathbf{X}_s)|\mathbb{1}_{[0, \tau(\mathbf{x}))}(s)$, so by dominated convergence

$$\phi(\mathbf{T}(\mathbf{x})) - \phi(\mathbf{x}) = \int_0^{\tau(\mathbf{x})} f(\mathbf{X}_s) ds \quad \text{for } \mu_0\text{-a.e. } \mathbf{x}.$$

Both sides are integrable in $\mathbf{x} \sim \mu_0$ (ϕ is bounded; the right side by Eq. (89)). Taking expectations, $\mathbb{E}_{\mu_0} \phi(\mathbf{T}) = \int \phi d(\mathbf{T}_{\#} \mu_0) = \int \phi d\mu_1$ by Theorem 3.8(iii), and the right side has expectation $\int f d\nu$ by Eq. (89). Hence $\int \phi d\mu_1 - \int \phi d\mu_0 = \int \nabla \phi \cdot \mathbf{b} d\nu$.

Right equality. The function $\nabla \phi$ is bounded and continuous and $\mathbb{E} \|\dot{\mathbf{I}}_t\| < \infty$ (Lemma B.3(i)), so $\mathbb{E} |\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t| < \infty$, and Lemma D.2 with $\mathbf{c} = 0$, $\mathbf{G} = \nabla \phi^\top$ and the tower property give

$$\mathbb{E} [\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t] = \mathbb{E} [\nabla \phi(\mathbf{I}_t) \cdot \mathbf{b}(\mathbf{I}_t)] = \int \nabla \phi \cdot \mathbf{b} q d\mathbf{x}, \quad (92)$$

the conditioning being on $\mathbf{I}_t$ under the law with $t \sim \mathcal{U}[0, 1]$, and the last step because q is the density of $\mathbf{I}_t$. Now fix $(\mathbf{x}_0, \mathbf{x}_1)$. Each $\alpha_k(t) = (1-t)^{a_k}$ is continuous on $[0, 1]$ and C^1 on $[0, 1)$, with $\int_0^1 |\dot{\alpha}_k| dt = 1$ (Lemma B.3(i)); hence $t \mapsto \mathbf{I}_t$ is continuous on $[0, 1]$, C^1 on $[0, 1)$, with $\mathbf{I}_0 = \mathbf{x}_0$, $\mathbf{I}_1 = \mathbf{x}_1$, and $\|\dot{\mathbf{I}}_t\| \leq \sum_{\ell, k} |\dot{\alpha}_k(t)| \|\mathbf{x}_0^{\ell, k} - \mathbf{x}_1^{\ell, k}\|$ is integrable on $[0, 1)$. The function $t \mapsto \phi(\mathbf{I}_t)$ is therefore C^1 on $[0, 1)$ with derivative $\nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t$; the fundamental theorem of calculus on $[0, t']$, $t' < 1$, followed by $t' \uparrow 1$ (continuity of $\phi(\mathbf{I}_t)$ at 1 on the left, dominated convergence with dominant $\|\nabla \phi\|_\infty \|\dot{\mathbf{I}}_t\|$ on the right) gives

$$\int_0^1 \nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t dt = \phi(\mathbf{x}_1) - \phi(\mathbf{x}_0).$$

Since t is uniform and independent of $(\mathbf{x}_0, \mathbf{x}_1)$, the left side of Eq. (92) is $\mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1} \int_0^1 \nabla \phi(\mathbf{I}_t) \cdot \dot{\mathbf{I}}_t dt$ by Fubini (justified by $\mathbb{E} \|\dot{\mathbf{I}}_t\| < \infty$), which equals $\mathbb{E} \phi(\mathbf{x}_1) - \mathbb{E} \phi(\mathbf{x}_0) = \int \phi d\mu_1 - \int \phi d\mu_0$. $\square$

So ν and $q d\mathbf{x}$ solve the same stationary continuity equation $\nabla \cdot (\mathbf{b} \cdot) = \mu_0 - \mu_1$. We do not need them to coincide: Theorem 3.8(iv), written with Eq. (27) and Tonelli as $\int \phi d\nu = \mathbb{E}_{\mu_0} \int_0^\tau \phi(\mathbf{X}_s) ds$, gives the one-sided comparison

$$\int \phi d\nu \leq \int \phi q d\mathbf{x} \quad \text{for every Borel } \phi \geq 0. \quad (93)$$

With $\phi = \mathbb{1}_\Omega$, $\mathbb{E}_{\mu_0} \tau = \nu(\Omega) \leq \int_\Omega q d\mathbf{x} \leq 1$. With $\phi = \|\mathbf{b}\| \mathbb{1}_\Omega$ and the conditional Jensen inequality $\|\mathbf{b}(\mathbf{I}_t)\| = \|\mathbb{E}[\dot{\mathbf{I}}_t | \mathbf{I}_t]\| \leq \mathbb{E}[\|\dot{\mathbf{I}}_t\| | \mathbf{I}_t]$,

$$\int \|\mathbf{b}\| d\nu \leq \int \|\mathbf{b}\| q d\mathbf{x} = \mathbb{E} \|\mathbf{b}(\mathbf{I}_t)\| \leq \mathbb{E} \|\dot{\mathbf{I}}_t\| < \infty$$

by Lemma B.3(i), so the hypothesis of Lemma E.1 holds.

Remark E.2 (Numerical check). We integrated the exact field (Proposition 3.3) from 1,000 prior samples with a midpoint rule ($ds = 2 \cdot 10^{-3}$, stopping within 10^{-3} of a vertex) and compared ν with q ; numerically Eq. (93) holds with near equality. The mean hitting time is $\mathbb{E}_{\mu_0} \tau = 1.000 \pm 0.014$ for flat DBTM on Δ^2 ($\sigma = 0.6$), 1.005 ± 0.004 on $\Delta^2 \times \Delta^1$ with $a = (5, 1)$, $\sigma = 0.55$, and 1.004 ± 0.010 with $a = (1, 1)$, $\sigma = (0.3, 0.8)$; the integrals of three test functions (distance to the data, norm, distance to the affine hull) under ν and under q agree to within 1.2% in all three geometries. In one dimension the identity follows directly from Eq. (90), since there $qb = \nu b$ is the flux $\mu_0((-\infty, x]) - \mu_1((-\infty, x])$.

E.1 Proof of Theorem 5.2

Proof. Idea. For (i) we compare $\mathbf{T}_\theta$ at the two ends of one exact trajectory. The exact map $\mathbf{T}$ is constant along it, while $\mathbf{T}_\theta$ changes at the rate $\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}$; so $\mathbf{T}_\theta(\mathbf{x})$ differs from $\mathbf{T}_\theta(\mathbf{T}(\mathbf{x}))$, which is within the boundary error of $\mathbf{T}(\mathbf{x})$, by the integral of the residual along the trajectory, and averaging over $\mathbf{x} \sim \mu_0$ turns that integral into an integral against ν . Part (ii) then follows from the comparison Eq. (93) of ν with q and from the coupling $(\mathbf{T}_\theta(\mathbf{x}), \mathbf{T}(\mathbf{x}))$ in the definition of W_1 . Part (iii) is Lemma D.1 once more: the excess of the loss over its one-step minimum is the squared distance between $\mathbf{T}_\theta$ and the regression step of Proposition 5.1, i.e. $\mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\|^2$. Part (iv) compares the distance to a vertex with the cross-entropy through $1 - p \leq -\log p$.

Write $\mathbf{r} := \mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}$ on Ω , which is Borel ($\mathbf{T}_\theta \in C^1(\Omega)$, $\mathbf{b} \in C^\infty(\Omega)$).

(i) If $\int \|\mathbf{r}\| d\nu = \infty$ there is nothing to prove. Otherwise, by Eq. (89) with $f = \|\mathbf{r}\|$ and Theorem 3.8(i)–(ii), for μ_0 -a.e. $\mathbf{x}$ we have $\mathbf{x} \in \mathcal{A}$ and $\int_0^{\tau(\mathbf{x})} \|\mathbf{r}(\mathbf{X}_s)\| ds < \infty$; fix such an $\mathbf{x}$. For $s < \tau(\mathbf{x})$, $\mathbf{X}_s \in \Omega$, where $\mathbf{T}_\theta$

is C^1 , and $s \mapsto \mathbf{X}_s$ is C^1 , so by the chain rule

$$\frac{d}{ds} \mathbf{T}_\theta(\mathbf{X}_s) = \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{X}_s) \dot{\mathbf{X}}_s = \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{X}_s) \mathbf{b}(\mathbf{X}_s) = \mathbf{r}(\mathbf{X}_s),$$

and the fundamental theorem of calculus, applied coordinatewise on $[0, s']$, gives $\mathbf{T}_\theta(\mathbf{X}_{s'}) - \mathbf{T}_\theta(\mathbf{x}) = \int_0^{s'} \mathbf{r}(\mathbf{X}_s) ds$ for $s' < \tau(\mathbf{x})$. Let $s' \uparrow \tau(\mathbf{x})$: on the left $\mathbf{X}_{s'} \rightarrow \mathbf{T}(\mathbf{x})$ and $\mathbf{T}_\theta$ is continuous on all of $\mathbb{R}^D$, in particular at $\mathbf{T}(\mathbf{x}) \in \mathcal{M}_1^+$; on the right, dominated convergence with dominant $\|\mathbf{r}(\mathbf{X}_s)\| \mathbb{1}_{[0, \tau(\mathbf{x}))}(s)$ applies. Thus

$$\mathbf{T}_\theta(\mathbf{T}(\mathbf{x})) - \mathbf{T}_\theta(\mathbf{x}) = \int_0^{\tau(\mathbf{x})} \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{X}_s) \mathbf{b}(\mathbf{X}_s) ds. \quad (94)$$

Adding and subtracting $\mathbf{T}(\mathbf{x})$,

$$\mathbf{T}_\theta(\mathbf{x}) - \mathbf{T}(\mathbf{x}) = [\mathbf{T}_\theta(\mathbf{T}(\mathbf{x})) - \mathbf{T}(\mathbf{x})] - \int_0^{\tau(\mathbf{x})} \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{X}_s) \mathbf{b}(\mathbf{X}_s) ds.$$

For an integrable $\mathbf{v} : [0, \tau] \rightarrow \mathbb{R}^D$ with $\int \mathbf{v} \neq \mathbf{0}$, the unit vector $\mathbf{e} := \int \mathbf{v} / \|\int \mathbf{v}\|$ gives $\|\int \mathbf{v}\| = \int \langle \mathbf{e}, \mathbf{v} \rangle \leq \int \|\mathbf{v}\|$ by Cauchy–Schwarz. With the triangle inequality,

$$\|\mathbf{T}_\theta(\mathbf{x}) - \mathbf{T}(\mathbf{x})\| \leq \|\mathbf{T}_\theta(\mathbf{T}(\mathbf{x})) - \mathbf{T}(\mathbf{x})\| + \int_0^{\tau(\mathbf{x})} \|\mathbf{r}(\mathbf{X}_s)\| ds \quad \text{for } \mu_0\text{-a.e. } \mathbf{x}. \quad (95)$$

All three terms are nonnegative Borel functions of $\mathbf{x}$ ($\mathbf{T}$ is Borel on $\mathcal{A}$, being locally constant). Take the expectation over $\mathbf{x} \sim \mu_0$. Since $\mathbf{T}_\# \mu_0 = \mu_1$ (Theorem 3.8(iii)), the change-of-variables formula for image measures with $g(\mathbf{y}) := \|\mathbf{T}_\theta(\mathbf{y}) - \mathbf{y}\| \geq 0$ gives

$$\mathbb{E}_{\mathbf{x} \sim \mu_0} \|\mathbf{T}_\theta(\mathbf{T}(\mathbf{x})) - \mathbf{T}(\mathbf{x})\| = \mathbb{E}_{\mathbf{x} \sim \mu_0} g(\mathbf{T}(\mathbf{x})) = \int g d(\mathbf{T}_\# \mu_0) = \mathbb{E}_{\mathbf{y} \sim \mu_1} \|\mathbf{T}_\theta(\mathbf{y}) - \mathbf{y}\| = \varepsilon_{\text{bnd}}(\theta),$$

and the last term has expectation $\int \|\mathbf{r}\| d\nu$ by Eq. (89). This proves (i).

(ii) *Second inequality.* Apply Eq. (93) to the Borel function $\phi = \|\mathbf{r}\| \mathbb{1}_\Omega \geq 0$:

$$\int \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\| d\nu \leq \int_\Omega \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\| q d\mathbf{x} = \mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\| = R(\theta),$$

where $\Omega^c = \mathcal{M}_1^+$ is finite, hence Lebesgue-null. Combined with (i), $\mathbb{E}_{\mu_0} \|\mathbf{T}_\theta - \mathbf{T}\| \leq \varepsilon_{\text{bnd}}(\theta) + R(\theta)$.

First inequality. $W_1(P, P') = \inf \mathbb{E} \|\mathbf{U} - \mathbf{U}'\|$, the infimum over all pairs of random vectors with $\mathbf{U} \sim P$ and $\mathbf{U}' \sim P'$. Let $\mathbf{x} \sim \mu_0$, $\mathbf{U} := \mathbf{T}_\theta(\mathbf{x})$ and $\mathbf{U}' := \mathbf{T}(\mathbf{x})$. Then $\mathbf{U} \sim (\mathbf{T}_\theta)_\# \mu_0$ by definition of the image measure and $\mathbf{U}' \sim \mathbf{T}_\# \mu_0 = \mu_1$ by Theorem 3.8(iii), so this pair is admissible in the infimum and $W_1((\mathbf{T}_\theta)_\# \mu_0, \mu_1) \leq \mathbb{E} \|\mathbf{U} - \mathbf{U}'\| = \mathbb{E}_{\mu_0} \|\mathbf{T}_\theta - \mathbf{T}\|$.

(iii) *The minimisation class.* We minimise $\mathcal{E}(\cdot; \mathbf{T}_\theta)$ over measurable $\mathbf{T}$ with $\mathbb{E} \|\mathbf{T}(\mathbf{I}_t) - \mathbf{T}_\theta(\mathbf{I}_t)\|^2 < \infty$, a class that contains $\mathbf{T}_\theta$; when $\mathbb{E} \|\mathbf{T}_\theta(\mathbf{I}_t)\|^2 < \infty$ it is the class $\{\mathbb{E} \|\mathbf{T}(\mathbf{I}_t)\|^2 < \infty\}$ of Proposition 5.1. Writing $\mathbf{T} = \mathbf{T}_\theta + \mathbf{S}$, the class is $\{\mathbb{E} \|\mathbf{S}(\mathbf{I}_t)\|^2 < \infty\}$, and

$$\mathcal{E}(\mathbf{T}_\theta + \mathbf{S}; \mathbf{T}_\theta) = \mathbb{E} \|\mathbf{S}(\mathbf{I}_t) - \mathbf{Y}\|^2, \quad \mathbf{Y} := \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \dot{\mathbf{I}}_t,$$

with $\mathbb{E} \|\mathbf{Y}\|^2 < \infty$ by hypothesis ($\mathbf{J}_{\mathbf{T}_\theta}$ is defined on Ω , and $\mathbf{I}_t \in \Omega$ almost surely).

Conditional mean. $\mathbb{E} \|\mathbf{Y}\| \leq (\mathbb{E} \|\mathbf{Y}\|^2)^{1/2} < \infty$, so Lemma D.2 with $\mathbf{c} = \mathbf{0}$ and $\mathbf{G} = \mathbf{J}_{\mathbf{T}_\theta}$ (extended by $\mathbf{0}$ on the null set $\mathcal{M}_1^+$) gives $\mathbf{m}(\mathbf{I}_t) := \mathbb{E}[\mathbf{Y} \mid \mathbf{I}_t] = \mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \mathbf{b}(\mathbf{I}_t) = \mathbf{r}(\mathbf{I}_t)$.

Equality. Lemma D.1 with $\mathbf{z} = \mathbf{I}_t$ gives, for every admissible $\mathbf{S}$,

$$\mathcal{E}(\mathbf{T}_\theta + \mathbf{S}; \mathbf{T}_\theta) = \mathbb{E} \|\mathbf{S}(\mathbf{I}_t) - \mathbf{r}(\mathbf{I}_t)\|^2 + \mathbb{E} \|\mathbf{r}(\mathbf{I}_t) - \mathbf{Y}\|^2,$$

so $\min_{\mathbf{T}} \mathcal{E}(\mathbf{T}; \mathbf{T}_\theta) = \mathbb{E} \|\mathbf{r}(\mathbf{I}_t) - \mathbf{Y}\|^2$, attained at $\mathbf{S} = \mathbf{r}$, i.e. at the regression step $\mathbf{T} = \mathbf{T}_\theta + \mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}$ of Proposition 5.1. Taking $\mathbf{S} = \mathbf{0}$, i.e. $\mathbf{T} = \mathbf{T}_\theta$,

$$\mathcal{E}(\mathbf{T}_\theta; \mathbf{T}_\theta) = \mathbb{E} \|\mathbf{r}(\mathbf{I}_t)\|^2 + \min_{\mathbf{T}} \mathcal{E}(\mathbf{T}; \mathbf{T}_\theta) = \mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\|^2 + \min_{\mathbf{T}} \mathcal{E}(\mathbf{T}; \mathbf{T}_\theta),$$

which is the equality in Eq. (29).

First inequality. $q \, d\mathbf{x}$ is a probability measure, so by the Cauchy–Schwarz inequality $R(\theta) = \mathbb{E}_q[1 \cdot \|\mathbf{r}\|] \leq (\mathbb{E}_q 1)^{1/2} (\mathbb{E}_q \|\mathbf{r}\|^2)^{1/2}$, i.e. $R(\theta)^2 \leq \mathbb{E}_q \|\mathbf{r}\|^2$.

Last inequality. $\mathcal{E}(\mathbf{T}_\theta; \mathbf{T}_\theta) = \mathbb{E} \|\mathbf{T}_\theta(\mathbf{I}_t) - \mathbf{T}_\theta(\mathbf{I}_t) - \mathbf{Y}\|^2 = \mathbb{E} \|\mathbf{Y}\|^2$ and $\min_{\mathbf{T}} \mathcal{E} \geq 0$, so $\mathbb{E}_q \|\mathbf{r}\|^2 = \mathbb{E} \|\mathbf{Y}\|^2 - \min_{\mathbf{T}} \mathcal{E} \leq \mathbb{E} \|\mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2$.

Transport loss. The stop-gradient does not change values, so the summand of Eq. (25) at θ is $\|\mathbf{T}_\theta^{\ell,k} - \mathbf{J}_{\mathbf{T}_\theta^{\ell,k}} \dot{\mathbf{I}}_t\|^2 = \|\mathbf{J}_{\mathbf{T}_\theta^{\ell,k}}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2$. As $\mathbf{J}_{\mathbf{T}_\theta^{\ell,k}} \dot{\mathbf{I}}_t$ is block (ℓ, k) of $\mathbf{J}_{\mathbf{T}_\theta} \dot{\mathbf{I}}_t$, the sum over blocks is $\|\mathbf{J}_{\mathbf{T}_\theta}(\mathbf{I}_t) \dot{\mathbf{I}}_t\|^2$, and the expectation is the right-hand side of Eq. (29).

(iv) Let $\mathbf{y} = c \mathbf{e}_j \in \text{supp } \mu_1$ (vertices scaled by c), fix a block (ℓ, k) with true symbol $i = j_{\ell,k}$, so $\mathbf{y}^{\ell,k} = c \mathbf{e}_i$, and write $\mathbf{T}_\theta^{\ell,k}(\mathbf{y}) = c \mathbf{p}$ with $\mathbf{p} = \text{softmax}(\cdot)$ in the open simplex, so $p_i \in (0, 1]$. Then $\text{CE}(\mathbf{T}_\theta^{\ell,k}(\mathbf{y})/c, \mathbf{y}^{\ell,k}/c) = \text{CE}(\mathbf{p}, \mathbf{e}_i) = -\log p_i$, and

$$\|c\mathbf{p} - c\mathbf{e}_i\|_2 \leq c\|\mathbf{p} - \mathbf{e}_i\|_1 = c\left((1 - p_i) + \sum_{i' \neq i} p_{i'}\right) = 2c(1 - p_i) \leq 2c(-\log p_i) = 2c \text{CE}(\mathbf{p}, \mathbf{e}_i),$$

using $\|\cdot\|_2 \leq \|\cdot\|_1$, $\sum_{i'} p_{i'} = 1$, and $\log p \leq p - 1$ (concavity of log) for $p \in (0, 1]$. Since $\|\mathbf{v}\|_2 = (\sum_{\ell,k} \|\mathbf{v}^{\ell,k}\|_2^2)^{1/2} \leq \sum_{\ell,k} \|\mathbf{v}^{\ell,k}\|_2$,

$$\|\mathbf{T}_\theta(\mathbf{y}) - \mathbf{y}\|_2 \leq \sum_{\ell,k} \|\mathbf{T}_\theta^{\ell,k}(\mathbf{y}) - \mathbf{y}^{\ell,k}\|_2 \leq 2c \sum_{\ell,k} \text{CE}(\mathbf{T}_\theta^{\ell,k}(\mathbf{y})/c, \mathbf{y}^{\ell,k}/c),$$

and taking expectations over $\mathbf{y} = \mathbf{x}_1 \sim \mu_1$ gives the claim. $\square$

E.2 Proof of Corollary 5.3

Proof. Idea. Theorem 5.2 bounds a Euclidean distance, while decoding takes an argmax, so we need a margin: a block can decode to a wrong symbol only if the learned block is at distance at least $c/\sqrt{2}$ from the scaled vertex, and Markov’s inequality turns the expected distance into a probability. Parts (ii) and (iii) are then counting. A wrong token needs a wrong block, the union bound and the Cauchy–Schwarz inequality collect the blocks into $\mathbb{E} \|\mathbf{T}_\theta(\mathbf{x}_0) - \mathbf{T}(\mathbf{x}_0)\|$, and two random variables defined from the same $\mathbf{x}_0$ have laws within total variation $\mathbb{P}(\hat{\mathbf{W}} \neq \mathbf{W})$.

Write $\Delta := \mathbf{T}_\theta(\mathbf{x}_0) - \mathbf{T}(\mathbf{x}_0)$, with blocks $\Delta^{\ell,k}$ and position vectors $\Delta^\ell = (\Delta^{\ell,k})_k$. By Theorem 3.8, almost surely $\mathbf{T}(\mathbf{x}_0) \in \mathcal{M}_1^+$ (vertices scaled by c).

(i) Fix a block and let $\mathbf{T}_\theta^{\ell,k}(\mathbf{x}_0) = c \mathbf{e}_j$. Since $c > 0$, j is the unique argmax of $c \mathbf{e}_j$, so $\text{dec}(\mathbf{T}(\mathbf{x}_0))^{\ell,k} = j$ regardless of the tie rule. Let $\mathbf{p} := \mathbf{T}_\theta^{\ell,k}(\mathbf{x}_0) \in \mathbb{R}^{v_k}$ and suppose $\text{dec}(\mathbf{T}_\theta(\mathbf{x}_0))^{\ell,k} = i \neq j$; then $p_i = \max_{i'} p_{i'} \geq p_j$. Keeping only coordinates i and j ,

$$\|\mathbf{p} - c\mathbf{e}_j\|^2 \geq (p_j - c)^2 + p_i^2.$$

If $p_j \geq 0$, then $p_i \geq p_j \geq 0$ gives $p_i^2 \geq p_j^2$, and with $y = p_j$,

$$(p_j - c)^2 + p_i^2 \geq (y - c)^2 + y^2 = 2\left(y - \frac{c}{2}\right)^2 + \frac{c^2}{2} \geq \frac{c^2}{2}.$$

If $p_j < 0$, then $(p_j - c)^2 = (c + |p_j|)^2 > c^2 > c^2/2$. Either way $\|\Delta^{\ell,k}\| \geq c/\sqrt{2}$, so by Markov’s inequality

$$\mathbb{P}(\text{dec}(\mathbf{T}_\theta(\mathbf{x}_0))^{\ell,k} \neq \text{dec}(\mathbf{T}(\mathbf{x}_0))^{\ell,k}) \leq \mathbb{P}(\|\Delta^{\ell,k}\| \geq c/\sqrt{2}) \leq \frac{\sqrt{2}}{c} \mathbb{E} \|\Delta^{\ell,k}\|.$$

(ii) The token at position ℓ is γ^{-1} of the tuple of its K block decodings. If all K blocks of position ℓ agree with those of $\text{dec}(\mathbf{T}(\mathbf{x}_0))$, the tuple equals that of the exact sample, which lies in the image of γ because $\mathbf{T}(\mathbf{x}_0) \in \text{supp } \mu_1$, so the token is decoded correctly. Hence a wrong token at position ℓ requires a wrong block, and by the union bound and (i),

$$\mathbb{P}(\text{token } \ell \text{ wrong}) \leq \sum_{k=1}^K \frac{\sqrt{2}}{c} \mathbb{E} \|\Delta^{\ell,k}\| \leq \frac{\sqrt{2}}{c} \mathbb{E} \left[\sqrt{K} \left(\sum_{k=1}^K \|\Delta^{\ell,k}\|^2 \right)^{1/2} \right] = \frac{\sqrt{2K}}{c} \mathbb{E} \|\Delta^\ell\|,$$

by the Cauchy–Schwarz inequality $\sum_{k=1}^K a_k \leq \sqrt{K}(\sum_k a_k^2)^{1/2}$ applied pointwise. The expected number of wrong tokens is $\sum_\ell \mathbb{P}(\text{token } \ell \text{ wrong})$ by linearity. Summing, and applying Cauchy–Schwarz over the L positions pointwise, $\sum_\ell \|\Delta^\ell\| \leq \sqrt{L}(\sum_\ell \|\Delta^\ell\|^2)^{1/2} = \sqrt{L}\|\Delta\|$, so

$$\mathbb{E}[\#\text{wrong tokens}] \leq \frac{\sqrt{2K}}{c} \sum_\ell \mathbb{E}\|\Delta^\ell\| \leq \frac{\sqrt{2KL}}{c} \mathbb{E}\|\Delta\| \leq \frac{\sqrt{2KL}}{c}(\varepsilon_{\text{bnd}} + R)$$

by Eq. (28).

(iii) Let $\mathbf{W} := \gamma^{-1}(\text{dec}(\mathbf{T}(\mathbf{x}_0)))$ and $\hat{\mathbf{W}} := \gamma^{-1}(\text{dec}(\mathbf{T}_\theta(\mathbf{x}_0)))$ (with an extra value “fail” where γ^{-1} is undefined). On $\mathcal{M}_1^+$ the decoding inverts the embedding of Definition 3.1, and μ_1 is the image of the data distribution under that embedding; since $\mathbf{T}(\mathbf{x}_0) \sim \mathbf{T}_\# \mu_0 = \mu_1$, $\mathbf{W}$ has the data law. For every set A of outcomes,

$$\mathbb{P}(\hat{\mathbf{W}} \in A) - \mathbb{P}(\mathbf{W} \in A) \leq \mathbb{P}(\hat{\mathbf{W}} \in A, \mathbf{W} \notin A) \leq \mathbb{P}(\hat{\mathbf{W}} \neq \mathbf{W}),$$

and the same with the roles exchanged, so the total variation distance between the two laws is at most $\mathbb{P}(\hat{\mathbf{W}} \neq \mathbf{W})$. Since $\mathcal{K}[\hat{\mathbf{W}} \neq \mathbf{W}] \leq \#\text{wrong tokens}$, (ii) bounds this probability by $\frac{\sqrt{2KL}}{c}(\varepsilon_{\text{bnd}} + R)$, and a total variation distance is at most 1. $\square$

F Proofs for Section 6

F.1 Proof of Theorem 6.1

Proof. Idea. The log-odds of the truth is $\Lambda = \rho g_\star + \rho^2 - \log S$, where $S = \sum_{j \neq i^\star} e^{\rho g_j}$ is independent of $g_\star$. It therefore suffices to pin down $\log S$: Markov’s inequality on $\mathbb{E}S$ bounds it from above by about $\rho^2/2 + s^2/2$, and its largest term $e^{\rho G}$, with G the maximum of $n-1$ Gaussians, bounds it from below by about ρs . Near the threshold both bounds equal ρs up to lower-order terms, so $\Lambda \approx \rho(g_\star + x)$, and the only randomness left is the Gaussian $g_\star$ on the true coordinate. Passing Λ through the logistic function, which for large ρ is close to the indicator of $\{g_\star + x > 0\}$, gives $\Phi(x)$ for the mean; the limits come from a choice of w, y, η that makes all error terms vanish.

Let $m = n-1 \geq 2$ and $s = \sqrt{2 \log m} > 0$. We argue conditionally on $i^\star$; given $i^\star$, the coordinates $g_j = x_{0,j}/\sigma$, $j \in [n]$, are independent standard normal, and all bounds below hold conditionally, hence unconditionally. Use the notation of Eq. (61): $\pi_\star = 1/(1 + \Sigma)$ with $\Sigma = e^{-\rho g_\star - \rho^2} S$ and $S = \sum_{j \neq i^\star} e^{\rho g_j} > 0$. Since $\pi_\star/(1 - \pi_\star) = \frac{1/(1+\Sigma)}{\Sigma/(1+\Sigma)} = 1/\Sigma$,

$$\Lambda = \log \frac{\pi_\star}{1 - \pi_\star} = -\log \Sigma = \rho g_\star + \rho^2 - \log S. \quad (96)$$

The variable S depends only on $(g_j)_{j \neq i^\star}$ and is independent of $g_\star$.

Upper bound on S . For a standard normal g , $\mathbb{E}e^{\rho g} = e^{\rho^2/2}$ (proof of Lemma C.3), so by linearity $\mathbb{E}S = m e^{\rho^2/2} \in (0, \infty)$. Markov’s inequality applied to $S \geq 0$ gives $\mathbb{P}(S > e^y \mathbb{E}S) \leq e^{-y} \mathbb{E}S / \mathbb{E}S = e^{-y}$. On the complement $\{S \leq e^y \mathbb{E}S\}$, using $\log m = s^2/2$,

$$\log S \leq y + \log m + \frac{\rho^2}{2} = y + \frac{s^2}{2} + \frac{\rho^2}{2}. \quad (97)$$

Lower bound on S . Let $G = \max_{j \neq i^\star} g_j$. All terms of S are positive and one of them is $e^{\rho G}$, so $S \geq e^{\rho G}$. Put $x_w := s - w$, which is positive because $w < s$. By Lemma C.7 (with the m variables $(g_j)_{j \neq i^\star}$) and Lemma C.6,

$$\mathbb{P}(G < x_w) \leq \exp(-mQ(x_w)) \leq \exp\left(-m \frac{x_w}{1 + x_w^2} \varphi(x_w)\right) = e^{-\Lambda_n(w)},$$

where the last equality is the computation in the proof of Theorem 4.2(b) with $h = w$: $m\varphi(s-w) = \frac{m}{\sqrt{2\pi}} e^{-s^2/2} e^{sw-w^2/2} = \frac{1}{\sqrt{2\pi}} e^{sw-w^2/2}$, since $m e^{-s^2/2} = 1$. On the complement $\{G \geq s-w\}$, as $\rho > 0$,

$$\log S \geq \rho G \geq \rho(s-w). \quad (98)$$

Log-odds. Let $\mathcal{E} := \{S \leq e^y \mathbb{E}S\} \cap \{G \geq s - w\}$. It depends only on $(g_j)_{j \neq i^*}$, so it is independent of g_* , and by the union bound $\mathbb{P}(\mathcal{E}^c) \leq e^{-y} + e^{-\Lambda_n(w)} = \epsilon_n$. On $\mathcal{E}$, inserting Eq. (97) into Eq. (96) and using $\rho = s + x$, so that $\rho^2 - s^2 = 2sx + x^2$,

$$\Lambda \geq \rho g_* + \rho^2 - y - \frac{s^2}{2} - \frac{\rho^2}{2} = \rho g_* + \frac{\rho^2 - s^2}{2} - y = \rho g_* + sx + \frac{x^2}{2} - y,$$

and inserting Eq. (98) into Eq. (96), $\Lambda \leq \rho g_* + \rho^2 - \rho(s - w) = \rho g_* + \rho(\rho - s + w) = \rho g_* + \rho(x + w)$. This is Eq. (31), on an event of probability at least $1 - \epsilon_n$.

Elementary bounds for the logistic function. Let $\varsigma(u) = 1/(1+e^{-u})$. Then $\pi_* = 1/(1+\Sigma) = 1/(1+e^{-\Lambda}) = \varsigma(\Lambda)$ by Eq. (96), ς is strictly increasing with values in $(0, 1)$, and $\varsigma(u) + \varsigma(-u) = \frac{1}{1+e^{-u}} + \frac{e^{-u}}{e^{-u}+1} = 1$. For every $v \in \mathbb{R}$,

$$\varsigma(-v) = \frac{1}{1+e^v} \leq \frac{1}{e^v} = e^{-v}, \quad \varsigma(v) = 1 - \varsigma(-v) \geq 1 - e^{-v}. \quad (99)$$

Fix $\eta > 0$ and $u \in \mathbb{R}$; recall $\rho > 0$.

- If $u > -\eta$, then $\varsigma(\rho u) \leq 1 = \mathbb{K}\{u > -\eta\}$. If $u \leq -\eta$, then $\rho u \leq -\rho\eta$ and $\varsigma(\rho u) \leq \varsigma(-\rho\eta) \leq e^{-\rho\eta}$ by monotonicity and Eq. (99). In both cases

$$\varsigma(\rho u) \leq \mathbb{K}\{u > -\eta\} + e^{-\rho\eta}. \quad (100)$$

- If $u \geq \eta$, then $\varsigma(\rho u) \geq \varsigma(\rho\eta) \geq 1 - e^{-\rho\eta}$ by monotonicity and Eq. (99). If $u < \eta$, then $\varsigma(\rho u) > 0$. In both cases

$$\varsigma(\rho u) \geq \mathbb{K}\{u \geq \eta\}(1 - e^{-\rho\eta}). \quad (101)$$

Mean. Put $c := (x^2 + 2y)/(2\rho)$. Since $s = \rho - x$, $sx = \rho x - x^2$, and the lower bound in Eq. (31) equals

$$\rho g_* + \rho x - \frac{x^2}{2} - y = \rho g_* + \rho \left(x - \frac{x^2 + 2y}{2\rho} \right) = \rho(g_* + x - c).$$

Hence, as ς is increasing, on $\mathcal{E}$

$$\varsigma(\rho(g_* + x - c)) \leq \pi_* \leq \varsigma(\rho(g_* + x + w)). \quad (102)$$

Upper bound. Split $\mathbb{E}\pi_* = \mathbb{E}[\pi_* \mathbb{K}_{\mathcal{E}}] + \mathbb{E}[\pi_* \mathbb{K}_{\mathcal{E}^c}]$, use Eq. (102) and $\varsigma > 0$ in the first term and $\pi_* \leq 1$ in the second, then Eq. (100) with $u = g_* + x + w$, and finally the symmetry of $g_* \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}\pi_* \leq \mathbb{E}\varsigma(\rho(g_* + x + w)) + \mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}(g_* > -x - w - \eta) + e^{-\rho\eta} + \epsilon_n = \Phi(x + w + \eta) + e^{-\rho\eta} + \epsilon_n.$$

Lower bound. Drop $\mathbb{E}[\pi_* \mathbb{K}_{\mathcal{E}^c}] \geq 0$, use Eq. (102), the independence of g_* and $\mathcal{E}$, and Eq. (101) with $u = g_* + x - c$:

$$\mathbb{E}\pi_* \geq \mathbb{E}[\varsigma(\rho(g_* + x - c)) \mathbb{K}_{\mathcal{E}}] = \mathbb{E}\varsigma(\rho(g_* + x - c)) \mathbb{P}(\mathcal{E}) \geq \mathbb{P}(g_* \geq \eta - x + c)(1 - e^{-\rho\eta}) \mathbb{P}(\mathcal{E}).$$

Here $\mathbb{P}(g_* \geq \eta - x + c) = \Phi(x - c - \eta)$, and with $A := \Phi(x - c - \eta)(1 - e^{-\rho\eta}) \in [0, 1]$, $A \mathbb{P}(\mathcal{E}) \geq A(1 - \epsilon_n) \geq A - \epsilon_n$. This is Eq. (32).

Limits. Let $n \rightarrow \infty$, so $s \rightarrow \infty$, and take

$$w = \frac{2 \log s}{s}, \quad y = \sqrt{s}, \quad \eta = s^{-1/2}.$$

For $s > 1$ these are admissible: $y, \eta > 0$, $w > 0$, and $w < s$ because $2 \log s < 2(s - 1) < s^2$. Moreover $w \rightarrow 0$ and $\eta \rightarrow 0$.

(i) $\epsilon_n \rightarrow 0$. Since $sw = 2 \log s$, $e^{sw - w^2/2} = s^2 e^{-w^2/2}$, so $\Lambda_n(w) = \frac{s-w}{1+(s-w)^2} \frac{s^2 e^{-w^2/2}}{\sqrt{2\pi}}$. For s large, $w \leq 1$ and $s - w \geq s/2 \geq 1$; then $\frac{s-w}{1+(s-w)^2} \geq \frac{1}{2(s-w)} \geq \frac{1}{2s}$ (as $1 + v^2 \leq 2v^2$ for $v \geq 1$) and $e^{-w^2/2} \geq e^{-1/2}$, so $\Lambda_n(w) \geq \frac{s e^{-1/2}}{2\sqrt{2\pi}} \rightarrow \infty$. With $e^{-y} = e^{-\sqrt{s}} \rightarrow 0$ this gives $\epsilon_n \rightarrow 0$; the choices do not involve x , so the convergence is uniform in x .

(ii) *The remaining terms.* Fix $X > 0$ and let $|x| \leq X$. For $s > X$, $\rho = s + x \geq s - X > 0$, and

$$e^{-\rho\eta} \leq e^{-(s-X)/\sqrt{s}} \rightarrow 0, \quad 0 < c = \frac{x^2 + 2y}{2\rho} \leq \frac{X^2 + 2\sqrt{s}}{2(s-X)} \rightarrow 0, \quad w + \eta \rightarrow 0,$$

all uniformly in $|x| \leq X$. Let $\delta_n := \sup_{|x| \leq X} (w + \eta + c + e^{-\rho\eta}) \rightarrow 0$.

(iii) *Uniform convergence of the mean.* Φ is Lipschitz with constant $\sup \varphi = \varphi(0) = 1/\sqrt{2\pi} < 1$. From the upper bound in Eq. (32), $\mathbb{E}\pi_\star - \Phi(x) \leq \Phi(x + w + \eta) - \Phi(x) + e^{-\rho\eta} + \epsilon_n \leq (w + \eta) + e^{-\rho\eta} + \epsilon_n$. From the lower bound, using $\Phi \leq 1$ so that $\Phi(\cdot)(1 - e^{-\rho\eta}) \geq \Phi(\cdot) - e^{-\rho\eta}$, $\mathbb{E}\pi_\star - \Phi(x) \geq \Phi(x - c - \eta) - \Phi(x) - e^{-\rho\eta} - \epsilon_n \geq -(c + \eta) - e^{-\rho\eta} - \epsilon_n$. Hence $\sup_{|x| \leq X} |\mathbb{E}\pi_\star(s_n + x) - \Phi(x)| \leq \delta_n + \epsilon_n \rightarrow 0$.

(iv) *The log-odds.* Fix x . Dividing Eq. (31) by s and writing $\rho/s = 1 + x/s$, $\rho(x + w)/s = (s + x)(x + w)/s = x + w + x(x + w)/s$, and $sx/s = x$, we obtain on $\mathcal{E}$

$$\frac{xg_\star}{s} + \frac{x^2}{2s} - \frac{y}{s} \leq \frac{\Lambda}{s} - (x + g_\star) \leq \frac{xg_\star}{s} + w + \frac{x(x + w)}{s},$$

hence $\left| \frac{\Lambda}{s} - (x + g_\star) \right| \leq \frac{|x||g_\star|}{s} + r_n$ on $\mathcal{E}$, with the deterministic $r_n := \frac{x^2}{2s} + \frac{1}{\sqrt{s}} + w + \frac{|x|(|x| + w)}{s} \rightarrow 0$. For $\delta > 0$ and n so large that $r_n < \delta/2$,

$$\mathbb{P}\left(\left|\frac{\Lambda}{s} - (x + g_\star)\right| > \delta\right) \leq \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(|x||g_\star| > \delta s/2) \leq \epsilon_n + \mathbb{P}(|x||g_\star| > \delta s/2) \rightarrow 0,$$

because $g_\star \sim \mathcal{N}(0, 1)$ does not depend on n and $s \rightarrow \infty$. $\square$

F.2 Proof of Proposition 6.2

Proof. Idea. The coordinate ℓ is chosen so that power schedules act by translation: $\rho(t) \leq \rho_0$ is equivalent to $-a \log(1 - t) \leq \log(1 + \sigma\rho_0)$, and after one more logarithm the exponent a enters only through the shift $-\log a$. The law of ℓ is a direct computation of $\mathbb{P}(\ell \leq l)$. What remains is to show that the length $\lambda_s(\sigma) = \log h(\sigma(s - z))$, with $h(y) = \log(1 + qy)/\log(1 + y)$, decreases strictly from $\log q$ to 0; after the quotient rule, monotonicity of h reduces to monotonicity of $u \mapsto u \log u/(u - 1)$.

Law of ℓ . Almost surely $t \in (0, 1)$, where $\ell(t)$ is defined. For $l \in \mathbb{R}$, since $\log$ and $\exp$ are increasing and $1 - \exp(-e^l) \in (0, 1)$,

$$\mathbb{P}(\ell \leq l) = \mathbb{P}(-\log(1 - t) \leq e^l) = \mathbb{P}(1 - t \geq \exp(-e^l)) = \mathbb{P}(t \leq 1 - \exp(-e^l)) = 1 - \exp(-e^l) = G(l),$$

and $G'(l) = e^l \exp(-e^l) = \exp(l - e^l) = \varphi_G(l)$.

Window. By Corollary 4.4, $\rho(t) = ((1 - t)^{-a} - 1)/\sigma$ is strictly increasing on $[0, 1]$ with $\rho(0) = 0$. For $\rho_0 > 0$ and $t \in (0, 1)$, since $\sigma > 0$, $a > 0$ and $\log$ is increasing,

$$\rho(t) \leq \rho_0 \iff (1 - t)^{-a} \leq 1 + \sigma\rho_0 \iff -a \log(1 - t) \leq \log(1 + \sigma\rho_0) \iff \ell(t) \leq \log \frac{\log(1 + \sigma\rho_0)}{a},$$

where the last step divides by a and takes logarithms of the two positive numbers $-\log(1 - t)$ and $\log(1 + \sigma\rho_0)/a$. The same chain holds with strict inequalities. As $s - z > 0$, the window is $\{s - z \leq \rho(t) \leq s + z\} = \{\rho \leq s + z\} \setminus \{\rho < s - z\}$, and $t = 0$ is not in it; with $\rho_0 = s + z$ and $\rho_0 = s - z$ this is $\{\ell \in [c, c']\}$ with $c = \log(\log(1 + \sigma(s - z))/a)$ and $c' = \log(\log(1 + \sigma(s + z))/a) = c + \lambda_s(\sigma)$.

The length. Put $q = (s + z)/(s - z) > 1$ and $y = \sigma(s - z) > 0$, so that $\sigma(s + z) = qy$ and $\lambda_s(\sigma) = \log h(y)$ with

$$h(y) = \frac{\log(1 + qy)}{\log(1 + y)}, \quad y > 0.$$

Limits. As $y \rightarrow 0$, $\log(1 + v)/v \rightarrow 1$ gives $h(y) = q \cdot \frac{\log(1 + qy)}{qy} \cdot \frac{y}{\log(1 + y)} \rightarrow q$. As $y \rightarrow \infty$, $h(y) = \frac{\log y + \log(q + 1/y)}{\log y + \log(1 + 1/y)} \rightarrow 1$. *Monotonicity.* Let $\psi(u) = u \log u/(u - 1)$ on $(1, \infty)$. Then

$$\psi'(u) = \frac{(\log u + 1)(u - 1) - u \log u}{(u - 1)^2} = \frac{u - 1 - \log u}{(u - 1)^2} > 0,$$

because $\log u < u - 1$ for $u \neq 1$; so ψ is strictly increasing. Let $A = 1 + y$ and $B = 1 + qy$; then $1 < A < B$ and $B - 1 = q(A - 1)$, so, since $q(A - 1) > 0$,

$$qA \log A = q\psi(A)(A - 1) < q\psi(B)(A - 1) = \psi(B)(B - 1) = B \log B.$$

By the quotient rule,

$$h'(y) = \frac{\frac{q}{B} \log A - \frac{1}{A} \log B}{(\log A)^2} = \frac{qA \log A - B \log B}{AB (\log A)^2} < 0.$$

Hence h is continuous and strictly decreasing on $(0, \infty)$ with limits q at 0 and 1 at ∞ ; by the intermediate value theorem it is a bijection of $(0, \infty)$ onto $(1, q)$. Composing with the increasing bijection $\sigma \mapsto \sigma(s - z)$ of $(0, \infty)$ and the increasing bijection $\log : (1, q) \rightarrow (0, \log q) = (0, L_s)$, the map λ_s is a strictly decreasing bijection of $(0, \infty)$ onto $(0, L_s)$.

Realisability. Every power schedule $(a, \sigma > 0)$ gives $c \in \mathbb{R}$ and $\lambda = \lambda_s(\sigma) \in (0, L_s)$. Conversely, given $c \in \mathbb{R}$ and $\lambda \in (0, L_s)$, take $\sigma = \lambda_s^{-1}(\lambda) > 0$ and $a = \log(1 + \sigma(s - z)) e^{-c} > 0$; then $\log(\log(1 + \sigma(s - z)))/a = \log e^c = c$, and the window is $[c, c + \lambda]$. $\square$

F.3 Proof of Theorem 6.3

Proof. Idea. By Proposition 6.2 the window share is the G -mass of an interval of length $\lambda < L_s$, which is smaller than the G -mass of the interval of length L_s with the same left end, so it suffices to maximise $f(c) = G(c + L_s) - G(c)$ over the position c . Since $\log \varphi_G(l) = l - e^l$ is concave, $f'(c) = \varphi_G(c + L_s) - \varphi_G(c)$ changes sign exactly once, where the density takes equal values at the two ends of the interval, and this gives c^* and Eq. (33) in closed form. The remaining statements are estimates: $M(L)$ against L times the maximum $1/e$ of φ_G , a power series for L_s , and a direct maximisation over σ of the explicit window share of the linear schedule.

Let $L := L_s > 0$. Since $t \sim \mathcal{U}[0, 1]$, the window share is the probability that t lies in the window, which by Proposition 6.2 is $W(z) = \mathbb{P}(\ell \in [c, c + \lambda]) = G(c + \lambda) - G(c)$ with $c \in \mathbb{R}$ and $0 < \lambda < L$. Since $\varphi_G > 0$, $G(c + L) - G(c + \lambda) = \int_{c+\lambda}^{c+L} \varphi_G > 0$, so $W(z) < f(c) := G(c + L) - G(c) \leq \sup f = M(L)$.

Maximisation of f . f is differentiable with $f'(c) = \varphi_G(c + L) - \varphi_G(c)$. Because $\log$ is strictly increasing and $\log \varphi_G(l) = l - e^l$, $f'(c)$ has the sign of

$$\log \varphi_G(c + L) - \log \varphi_G(c) = (c + L - e^{c+L}) - (c - e^c) = L - e^c(e^L - 1).$$

As $e^L - 1 > 0$, the right-hand side is strictly decreasing in c and vanishes exactly at $c^* = \log(L/(e^L - 1))$. So $f' > 0$ on $(-\infty, c^*)$ and $f' < 0$ on (c^*, ∞) : f is strictly increasing on $(-\infty, c^*]$, strictly decreasing on $[c^*, \infty)$, and

$$\sup f = f(c^*) = (1 - \exp(-e^{c^*+L})) - (1 - \exp(-e^{c^*})) = \exp(-e^{c^*}) - \exp(-e^{c^*+L}),$$

with $e^{c^*} = L/(e^L - 1)$ and $e^{c^*+L} = Le^L/(e^L - 1)$. This is Eq. (33).

Approaching the bound. Fix $c = c^*$, i.e. by Proposition 6.2 $a = \log(1 + \sigma(s - z)) e^{-c^*} = (e^L - 1) \log(1 + \sigma(s - z))/L$, and let $\sigma \rightarrow 0$. Then $\lambda_s(\sigma) \uparrow L$ (Proposition 6.2: $h(y) \rightarrow q$ as $y \rightarrow 0$), and by continuity of G , $W(z) = G(c^* + \lambda_s(\sigma)) - G(c^*) \rightarrow f(c^*) = M(L)$.

Location. For $L > 0$, $e^L - 1 > L$ gives $e^{c^*} < 1$, and $Le^L > e^L - 1$ (equivalently $1 - e^{-L} < L$) gives $e^{c^*+L} > 1$. Hence $c^* < 0 < c^* + L$: the optimal interval contains $\ell = 0$. Since $(\log \varphi_G)'(l) = 1 - e^l$ is positive for $l < 0$ and negative for $l > 0$, $\ell = 0$ is the mode of φ_G , with $\varphi_G(0) = e^{-1}$; and $\ell(t) = 0 \iff -\log(1 - t) = 1 \iff t = 1 - 1/e$.

$M(L) \leq L/e$. For every c , $f(c) = \int_c^{c+L} \varphi_G \leq L \varphi_G(0) = L/e$.

Expansion of L_s . With $u := z/s \in (0, 1)$, $L_s = \log \frac{1+u}{1-u} = 2 \sum_{k \geq 0} \frac{u^{2k+1}}{2k+1}$, so

$$0 \leq L_s - \frac{2z}{s} \leq \frac{2}{3} \sum_{k \geq 1} u^{2k+1} = \frac{2}{3} \frac{u^3}{1 - u^2} \leq \frac{8}{9} \frac{z^3}{s^3} \quad (s \geq 2z),$$

i.e. $L_s = 2z/s + O(s^{-3})$, and $M(L_s) \leq L_s/e = 2z/(es) + O(s^{-3})$; with $s = s_n$ this is the stated fraction.

The linear schedule. For $a = 1$, [Corollary 4.4](#) gives $t(\rho) = 1 - (1 + \sigma\rho)^{-1} = \sigma\rho/(1 + \sigma\rho)$. With $A = s + z$, $B = s - z$ (so $A - B = 2z$, $AB = s^2 - z^2 > 0$),

$$W(\sigma) = \frac{\sigma A}{1 + \sigma A} - \frac{\sigma B}{1 + \sigma B} = \frac{\sigma A(1 + \sigma B) - \sigma B(1 + \sigma A)}{(1 + \sigma A)(1 + \sigma B)} = \frac{2\sigma z}{(1 + \sigma A)(1 + \sigma B)} > 0.$$

Then $\frac{d}{d\sigma} \log W = \frac{1}{\sigma} - \frac{A}{1 + \sigma A} - \frac{B}{1 + \sigma B}$, and multiplying by $\sigma(1 + \sigma A)(1 + \sigma B) > 0$ gives

$$(1 + \sigma A)(1 + \sigma B) - \sigma A(1 + \sigma B) - \sigma B(1 + \sigma A) = 1 - \sigma^2 AB.$$

So $\log W$ increases on $(0, (AB)^{-1/2}]$ and decreases on $[(AB)^{-1/2}, \infty)$, and the maximum is at $\sigma = (AB)^{-1/2} = (s^2 - z^2)^{-1/2}$. There

$$(1 + \sigma A)(1 + \sigma B) = 1 + \sigma(A + B) + \sigma^2 AB = 2 + \frac{A + B}{\sqrt{AB}} = \frac{(\sqrt{A} + \sqrt{B})^2}{\sqrt{AB}},$$

so $W = \frac{2z}{\sqrt{AB}} \cdot \frac{\sqrt{AB}}{(\sqrt{A} + \sqrt{B})^2} = \frac{2z}{(\sqrt{A} + \sqrt{B})^2}$. Large s . $(\sqrt{A} + \sqrt{B})^2 = 2s + 2s\sqrt{1 - u^2}$, and $1 - v \leq \sqrt{1 - u^2} \leq 1 - v/2$ for $v \in [0, 1]$ gives $4s - 2z^2/s \leq (\sqrt{A} + \sqrt{B})^2 \leq 4s - z^2/s$, hence

$$\frac{z}{2s} \leq W_{\text{lin}} \leq \frac{z}{2s - z^2/s} = \frac{z}{2s} + \frac{z^3}{2s^2(2s - z^2/s)}, \quad W_{\text{lin}} = \frac{z}{2s} + O(s^{-3}).$$

For $M(L_s)$: $\varphi'_G(l) = (1 - e^l)\varphi_G(l)$, and with $v = e^l$, $|\varphi'_G(l)| \leq ve^{-v} + v^2e^{-v} \leq e^{-1} + 4e^{-2} < 1$. Since $[c^*, c^* + L]$ contains 0, every l in it has $|l| \leq L$, so

$$\left| M(L) - \frac{L}{e} \right| = \left| \int_{c^*}^{c^* + L} (\varphi_G(l) - \varphi_G(0)) dl \right| \leq \int_{c^*}^{c^* + L} |l| dl \leq L^2.$$

With $L = L_s = 2z/s + O(s^{-3})$, $M(L_s) = 2z/(es) + O(s^{-2})$, and $W_{\text{lin}}/M(L_s) = \frac{e}{4} \cdot \frac{1 + O(s^{-2})}{1 + O(s^{-1})} \rightarrow \frac{e}{4}$. $\square$

F.4 Proof of [Corollary 6.4](#)

Proof. Idea. With σ fixed, [Proposition 6.2](#) says that the exponent only translates a window of fixed length λ in the coordinate ℓ , so maximising over a is the maximisation over c in the proof of [Theorem 6.3](#), with L_s replaced by λ . The optimal exponent is then read off from the relation between c and a .

By [Proposition 6.2](#), with σ fixed the window of the schedule $(1 - t)^a$ is $[c, c + \lambda]$ in the coordinate ℓ , with $\lambda = \lambda_{s_n}(\sigma)$ independent of a and $c = \log(\log(1 + \sigma(s_n - z))/a)$, and $a \mapsto c$ is a strictly decreasing bijection of $(0, \infty)$ onto $\mathbb{R}$. The window share is $f(c) = G(c + \lambda) - G(c)$. As in the proof of [Theorem 6.3](#) with L_s replaced by λ , $f'(c)$ has the sign of $\lambda - e^c(e^\lambda - 1)$, which is positive for $c < c^* := \log(\lambda/(e^\lambda - 1))$ and negative for $c > c^*$; so f has the unique maximiser c^* , with $f(c^*) = M(\lambda)$. Solving $c^* = \log(\log(1 + \sigma(s_n - z))/a)$ for a gives $a^* = \log(1 + \sigma(s_n - z))e^{-c^*} = \log(1 + \sigma(s_n - z))(e^\lambda - 1)/\lambda$, which is [Eq. \(34\)](#). Since $s_n - z < s_n < s_n + z$ and $\rho(t)$ is increasing, $t^* = t(s_n)$ lies in the window $\{t : |\rho(t) - s_n| \leq z\}$, whose image in ℓ is $[c^*, c^* + \lambda]$; and $e^{c^*} = \lambda/(e^\lambda - 1) < 1 < \lambda e^\lambda/(e^\lambda - 1) = e^{c^* + \lambda}$ (because $e^\lambda > 1 + \lambda$), so this interval contains $\ell = 0$, i.e. $t = 1 - 1/e$. The numerical values follow from [Eq. \(34\)](#) and [Eq. \(14\)](#). $\square$

F.5 Proof of [Proposition 6.5](#)

Proof. Idea. In the coordinate ℓ the exponents become positions c_k of intervals of fixed lengths λ_k , z -ordering says that consecutive intervals follow each other without overlap, and the problem is to place K such intervals on the line so that the smallest G -mass is as large as possible. For a target value m , the admissible positions of interval k form a set $C_k(m)$, which is an interval because f_k increases and then decreases. Placing each interval as far to the left as its own range and its predecessor allow leaves the most room for the intervals that follow; Claims 1 and 2 show that this greedy placement succeeds exactly when some placement does, so m^* is the largest m for which it succeeds. The upper bounds come from the best single window, $M(\lambda_k)$, and from the windows being disjoint subintervals of $[0, 1]$.

Reduction to the coordinate ℓ . By [Proposition 6.2](#), with σ_k fixed, level k 's window is $\{t : \ell(t) \in [c_k, c_k + \lambda_k]\}$ with $c_k = \log(\log(1 + \sigma_k(s_{v_k} - z))/a_k)$, and $a_k \mapsto c_k$ is a strictly decreasing bijection of $(0, \infty)$ onto $\mathbb{R}$ with inverse $a_k = \log(1 + \sigma_k(s_{v_k} - z))e^{-c_k}$. All levels share t , and ℓ is a strictly increasing bijection of $(0, 1)$ onto $\mathbb{R}$, so $t_k(s_{v_k} + z) \leq t_{k+1}(s_{v_{k+1}} - z)$ reads $c_k + \lambda_k \leq c_{k+1}$. As in the proof of [Theorem 6.3](#), $W_k = f_k(c_k)$ with $f_k(c) := G(c + \lambda_k) - G(c) > 0$. The problem is therefore to maximise $\min_k f_k(c_k)$ over the *ordered* sequences $(c_k) \in \mathbb{R}^K$, i.e. those with $c_k + \lambda_k \leq c_{k+1}$ for $k < K$. Only $m > 0$ matters, since every share is positive.

$C_k(m)$ is a compact interval for $m > 0$. The computation in the proof of [Theorem 6.3](#), with L replaced by λ_k , shows that f_k is strictly increasing on $(-\infty, c_k^*]$ and strictly decreasing on $[c_k^*, \infty)$, $c_k^* = \log(\lambda_k/(e^{\lambda_k} - 1))$. (Equivalently, $f_k = \varphi_G * \mathcal{K}_{[-\lambda_k, 0]}$ is a convolution of two log-concave functions on $\mathbb{R}$ — $\log \varphi_G(l) = l - e^l$ is concave — and is log-concave by the Prékopa–Leindler inequality.) Let $c < c''$ lie in $C_k(m)$ and $c \leq c''' \leq c''$. If $c''' \leq c_k^*$ then $f_k(c''') \geq f_k(c) \geq m$; otherwise $c_k^* < c''' \leq c''$ and $f_k(c''') \geq f_k(c'') \geq m$. So $C_k(m)$ is an interval. It is closed because f_k is continuous, and bounded because $f_k(c) \leq G(c + \lambda_k) \rightarrow 0$ as $c \rightarrow -\infty$ and $f_k(c) \leq 1 - G(c) \rightarrow 0$ as $c \rightarrow \infty$, while $m > 0$. We write $C_k(m) = [\alpha_k(m), \beta_k(m)]$ when it is nonempty (and use the conventions $\inf \emptyset = +\infty$, $\sup \emptyset = -\infty$, so the chain fails at an empty $C_k(m)$).

Feasibility. Call $m > 0$ *feasible* if some ordered (c'_k) has $f_k(c'_k) \geq m$ for all k , i.e. $c'_k \in C_k(m)$.

Claim 1: if m is feasible, the greedy chain exists, and $e_k \leq c'_k + \lambda_k$ for all k and every feasible witness (c'_k) . Induction on k . For $k = 1$: $c'_1 \in C_1(m)$, so $C_1(m) \neq \emptyset$ and $c_1 = \max\{-\infty, \alpha_1\} = \alpha_1 \leq c'_1 \leq \beta_1$; hence $c_1 \leq \sup C_1(m)$ and $e_1 = c_1 + \lambda_1 \leq c'_1 + \lambda_1$. Step $k - 1 \rightarrow k$: by the induction hypothesis and the order of (c'_j) , $e_{k-1} \leq c'_{k-1} + \lambda_{k-1} \leq c'_k$; also $\alpha_k \leq c'_k \leq \beta_k$ because $c'_k \in C_k(m)$. Therefore $c_k = \max\{e_{k-1}, \alpha_k\} \leq c'_k \leq \beta_k = \sup C_k(m)$, so the chain continues, and $e_k = c_k + \lambda_k \leq c'_k + \lambda_k$.

Claim 2: if the chain exists for m , then (c_k) is a feasible witness. For each k , $\alpha_k \leq c_k \leq \beta_k$, so $c_k \in C_k(m)$ and $f_k(c_k) \geq m$; and $c_{k+1} \geq e_k = c_k + \lambda_k$ is the order condition.

The optimum. Feasibility is monotone: if $0 < m' \leq m$ then $C_k(m) \subseteq C_k(m')$, so a witness for m is one for m' . The set of feasible m is nonempty: $c_k := k \max_j \lambda_j$ is ordered and gives $\min_k f_k(c_k) =: m_0 > 0$. On the set $\Gamma := \{(c_k) \text{ ordered} : c_k \in C_k(m_0) \forall k\}$, which is nonempty, closed and bounded, the continuous function $(c_k) \mapsto \min_k f_k(c_k)$ attains its maximum $m_1 \geq m_0$; every ordered sequence with $\min_k f_k(c_k) \geq m_0$ lies in Γ , so $m_1 = m^*$, and m^* is attained. Thus the feasible set is $(0, m^*]$, and by Claims 1 and 2 it coincides with the set of $m > 0$ for which the greedy chain exists. Hence m^* is the largest such m ; it can be located by bisection ([Table 1](#)), and the chain at $m = m^*$ gives, through $a_k = \log(1 + \sigma_k(s_{v_k} - z))e^{-c_k}$, z -ordered exponents with $\min_k W_k = m^*$.

Bounds. Each share satisfies $W_k = f_k(c_k) \leq \sup_c f_k(c) = M(\lambda_k)$, the supremum of [Theorem 6.3](#) with L_s replaced by λ_k ; so $m^* \leq \min_k M(\lambda_k)$. In t , the windows of a z -ordered choice are intervals in $[0, 1]$ that overlap at most in endpoints, so their lengths satisfy $\sum_k W_k \leq 1$, and $K \min_k W_k \leq \sum_k W_k \leq 1$ gives $m^* \leq 1/K$. $\square$

F.6 Proof of [Theorem 6.7](#)

Proof. Idea. Fano's inequality turns the error probability into a bound on the mutual information between i^* and what the reader sees. The reader sees $W\mathbf{I}_t$, which carries the same information as the projection of $\mathbf{I}_t$ onto an r -dimensional subspace, and there the observation is a Gaussian channel $\mathbf{Z} = \rho \mathbf{u}_{i^*} + \boldsymbol{\zeta}$ with unit noise and average signal power $\rho^2 r/n$, since the n projected vertices share the total squared length r . A Gaussian channel with r dimensions and this power carries at most $\frac{r}{2} \log(1 + \rho^2/n)$ nats, by the maximum-entropy property of the Gaussian and the AM–GM inequality applied to the eigenvalues of the covariance.

Logarithms are natural and information quantities are in nats. We use three standard facts.

- *Fano's inequality.* If V and $\hat{V}$ take values in a finite set of size $n \geq 2$ and $P_e = \mathbb{P}(\hat{V} \neq V)$, then $H(V | \hat{V}) \leq h_b(P_e) + P_e \log(n - 1)$, where $h_b(p) = -p \log p - (1 - p) \log(1 - p) \leq \log 2$.
- *Data processing.* If $V \rightarrow Y \rightarrow Y'$ is a Markov chain (Y' is conditionally independent of V given Y ; in particular if Y' is a measurable function of Y), then $I(V; Y') \leq I(V; Y)$.
- *Gaussian maximum entropy.* If a random vector $\mathbf{Z} \in \mathbb{R}^r$ has a density and a positive definite covariance matrix $\mathbf{K}$, then its differential entropy satisfies $h(\mathbf{Z}) \leq \frac{1}{2} \log((2\pi e)^r \det \mathbf{K})$, with equality for $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$.

Reduction to r coordinates. Let $\mathbf{U} \in \mathbb{R}^{r \times n}$ have orthonormal rows spanning the row space of W , so $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_r$ and $P := \mathbf{U}^\top \mathbf{U}$ is the orthogonal projection onto that row space. Each row of W lies in the

row space, so $W(\mathbf{I} - P) = 0$, i.e. $W = WP = W\mathbf{U}^\top\mathbf{U}$. The matrix $W\mathbf{U}^\top \in \mathbb{R}^{d \times r}$ has rank r , because $r = \text{rank } W = \text{rank}(W\mathbf{U}^\top\mathbf{U}) \leq \text{rank}(W\mathbf{U}^\top) \leq r$; hence it has a left inverse $\mathbf{L} \in \mathbb{R}^{r \times d}$, $\mathbf{L}W\mathbf{U}^\top = \mathbf{I}_r$. Therefore $W\mathbf{x} = (W\mathbf{U}^\top)\mathbf{U}\mathbf{x}$ and $\mathbf{U}\mathbf{x} = \mathbf{L}W\mathbf{x}$: the vectors $W\mathbf{x}$ and $\mathbf{U}\mathbf{x}$ are linear functions of each other. With $\mathbf{I}_t = (1 - \alpha)\mathbf{e}_{i^*} + \alpha\sigma\mathbf{g}$ and $\mathbf{Z} := \mathbf{U}\mathbf{I}_t/(\alpha\sigma)$, every estimator is $\hat{i} = \psi(W\mathbf{I}_t) = \psi(\alpha\sigma W\mathbf{U}^\top\mathbf{Z})$, a measurable function of $\mathbf{Z}$. With $\mathbf{u}_i = \mathbf{U}\mathbf{e}_i$ and $\boldsymbol{\zeta} = \mathbf{U}\mathbf{g}$,

$$\mathbf{Z} = \rho\mathbf{u}_{i^*} + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{U}^\top) = \mathcal{N}(\mathbf{0}, \mathbf{I}_r) \text{ independent of } i^*,$$

since $\rho = (1 - \alpha)/(\alpha\sigma)$ and $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is independent of i^* .

Fano and data processing. Let $P_e = \mathbb{P}(\hat{i} \neq i^*)$. As i^* is uniform, $H(i^*) = \log n$, and by Fano's inequality with $h_b \leq \log 2$ and $\log(n - 1) \leq \log n$,

$$I(i^*; \hat{i}) = H(i^*) - H(i^* | \hat{i}) \geq \log n - \log 2 - P_e \log n, \quad \text{i.e.} \quad P_e \geq 1 - \frac{I(i^*; \hat{i}) + \log 2}{\log n}.$$

Let $\mathbf{X} = \rho\mathbf{u}_{i^*}$. The chain $i^* \rightarrow \mathbf{X} \rightarrow \mathbf{Z} \rightarrow \hat{i}$ is Markov: $\mathbf{X}$ is a function of i^* , given $\mathbf{X}$ the vector $\mathbf{Z} = \mathbf{X} + \boldsymbol{\zeta}$ does not depend further on i^* because $\boldsymbol{\zeta}$ is independent of i^* , and $\hat{i}$ is a function of $\mathbf{Z}$. Data processing gives $I(i^*; \hat{i}) \leq I(i^*; \mathbf{Z}) \leq I(\mathbf{X}; \mathbf{Z})$.

Gaussian bound. $\mathbf{X}$ takes finitely many values, and given $\mathbf{X} = \boldsymbol{\xi}$ the vector $\mathbf{Z}$ has the density $\varphi_r(\cdot - \boldsymbol{\xi})$ of $\mathcal{N}(\boldsymbol{\xi}, \mathbf{I}_r)$; so $\mathbf{Z}$ has a bounded density, finite second moment, and $I(\mathbf{X}; \mathbf{Z}) = h(\mathbf{Z}) - h(\mathbf{Z} | \mathbf{X})$ with $h(\mathbf{Z} | \mathbf{X}) = \sum_{\boldsymbol{\xi}} \mathbb{P}(\mathbf{X} = \boldsymbol{\xi}) h(\boldsymbol{\xi} + \boldsymbol{\zeta}) = h(\boldsymbol{\zeta}) = \frac{1}{2} \log(2\pi e)^r$ by translation invariance. By independence of $\mathbf{X}$ and $\boldsymbol{\zeta}$, $\text{Cov}(\mathbf{Z}) = \text{Cov}(\mathbf{X}) + \mathbf{I}_r$, which is positive definite. The maximum-entropy bound gives

$$I(\mathbf{X}; \mathbf{Z}) \leq \frac{1}{2} \log((2\pi e)^r \det(\mathbf{I}_r + \text{Cov } \mathbf{X})) - \frac{1}{2} \log(2\pi e)^r = \frac{1}{2} \log \det(\mathbf{I}_r + \text{Cov } \mathbf{X}).$$

AM-GM. Let $\lambda_1, \dots, \lambda_r \geq 0$ be the eigenvalues of the positive semidefinite $\text{Cov } \mathbf{X}$. By the inequality of arithmetic and geometric means applied to the positive numbers $1 + \lambda_i$,

$$\det(\mathbf{I}_r + \text{Cov } \mathbf{X}) = \prod_{i=1}^r (1 + \lambda_i) \leq \left(\frac{1}{r} \sum_{i=1}^r (1 + \lambda_i) \right)^r = \left(1 + \frac{\text{tr Cov } \mathbf{X}}{r} \right)^r.$$

Moreover $\text{tr Cov } \mathbf{X} = \mathbb{E}\|\mathbf{X}\|^2 - \|\mathbb{E}\mathbf{X}\|^2 \leq \mathbb{E}\|\mathbf{X}\|^2$, and since i^* is uniform,

$$\mathbb{E}\|\mathbf{X}\|^2 = \frac{\rho^2}{n} \sum_{i=1}^n \|\mathbf{U}\mathbf{e}_i\|^2 = \frac{\rho^2}{n} \|\mathbf{U}\|_F^2 = \frac{\rho^2}{n} \text{tr}(\mathbf{U}\mathbf{U}^\top) = \frac{\rho^2 r}{n}.$$

As $\log$ is increasing, $I(\mathbf{X}; \mathbf{Z}) \leq \frac{r}{2} \log(1 + \rho^2/n)$. Combining the three displays gives [Eq. \(35\)](#).

Consequences. If $P_e \leq \varepsilon$, then [Eq. \(35\)](#) and $\log n > 0$ give $\frac{r}{2} \log(1 + \rho^2/n) + \log 2 \geq (1 - \varepsilon) \log n$, i.e. $\log(1 + \rho^2/n) \geq 2\kappa/r$ with $\kappa := (1 - \varepsilon) \log n - \log 2$. If $\kappa \leq 0$ the claimed bound is trivial. Otherwise $\rho^2/n \geq e^{2\kappa/r} - 1 \geq 2\kappa/r$ by $e^v - 1 \geq v$, i.e. $\rho^2 \geq \frac{2n}{r} \kappa$. For $r = n$ this reads $\rho^2 \geq 2\kappa$; directly, $\frac{n}{2} \log(1 + \rho^2/n) \leq \frac{\rho^2}{2}$ (by $\log(1 + v) \leq v$) turns [Eq. \(35\)](#) into $P_e \geq 1 - (\rho^2/2 + \log 2)/\log n$, which gives the same condition. $\square$

F.7 Proof of [Theorem 6.8](#)

Proof. Idea. A sample is recovered only if $f(W\mathbf{x})$ tells where the argmax J of $\mathbf{x}$ is. The projection reveals $\mathbf{m} = P\mathbf{x}$, in which coordinate j has variance π_j with $\sum_j \pi_j = r$, so at most r/θ coordinates are revealed with variance at least θ . Either the argmax sits on one of these, which has probability at most $r/(\theta n)$ because J is uniform, or the estimator names a poorly revealed coordinate; given $\mathbf{m}$, such a coordinate is a small m_j plus independent noise of variance at most one, and it rarely reaches the level τ that the maximum of n Gaussians exceeds with high probability. The proof bounds these cases separately and then chooses θ, c, τ so that all terms except $r/(\theta n)$ vanish.

Normalisation. Here $\mathbf{x}_0 \sim \mu_0 = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Replacing $\mathbf{x}_0$ by $\mathbf{x}_0/\sigma$ and f by $f(\sigma \cdot)$ (still measurable) leaves $\arg \max_i x_{0,i}$, $f(W\mathbf{x}_0)$, $\hat{i}(\mathbf{x}_0)$ and $\mathbf{T}(\mathbf{x}_0)$ unchanged, so we may take $\sigma = 1$: $\mathbf{x} := \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Let $J = \arg \max_i x_i$; ties have probability zero, so J is almost surely unique, and it is uniform on $[n]$ because the law of $\mathbf{x}$ is invariant under permutations of the coordinates. By [Theorem 4.9](#), $\mathbf{T}(\mathbf{x}) = \mathbf{e}_J$ almost surely.

The projected state. Let $\mathbf{U}$ and $P = \mathbf{U}^\top \mathbf{U}$ be as in the proof of [Theorem 6.7](#), and $\mathbf{m} := P\mathbf{x}$. Since $W = W\mathbf{U}^\top \mathbf{U} = W\mathbf{U}^\top \mathbf{U}P$ (as $\mathbf{U}P = \mathbf{U}\mathbf{U}^\top \mathbf{U} = \mathbf{U}$), $W\mathbf{x} = W\mathbf{U}^\top \mathbf{U}\mathbf{m}$, so $f(W\mathbf{x})$, and the nearest vertex $\hat{i}$ (ties broken by the smallest index), are measurable functions of $\mathbf{m}$: $\hat{i} = h(\mathbf{m})$. Let $\pi_j := P_{jj} = \|\mathbf{U}e_j\|^2$. As P is an orthogonal projection, $\pi_j = \|Pe_j\|^2 \in [0, 1]$ and $\sum_j \pi_j = \text{tr} P = \text{tr}(\mathbf{U}\mathbf{U}^\top) = r$. Fix $\theta \in (0, 1]$ and $0 \leq c \leq \tau$, and let $\mathcal{K} = \{j : \pi_j \geq \theta\}$; then $\theta|\mathcal{K}| \leq \sum_{j \in \mathcal{K}} \pi_j \leq r$, i.e. $|\mathcal{K}| \leq r/\theta$.

Decomposition. Since $\{\hat{i} = J\} \subseteq \{J \in \mathcal{K}\} \cup \{\hat{i} = J \notin \mathcal{K}\}$ and $\{\hat{i} = J \notin \mathcal{K}\} \subseteq \{\max_i x_i < \tau\} \cup \{x_i \geq \tau, \hat{i} \notin \mathcal{K}\}$ (on $\{\hat{i} = J\}$, $x_i = \max_i x_i$), the union bound gives

$$\mathbb{P}(\hat{i} = J) \leq \mathbb{P}(J \in \mathcal{K}) + \mathbb{P}(\max_i x_i < \tau) + \mathbb{P}(x_i \geq \tau, \hat{i} \notin \mathcal{K}). \quad (103)$$

As J is uniform, $\mathbb{P}(J \in \mathcal{K}) = |\mathcal{K}|/n \leq r/(\theta n)$, and [Lemma C.7](#) with the n variables $x_1, \dots, x_n$ gives $\mathbb{P}(\max_i x_i < \tau) \leq e^{-nQ(\tau)}$.

Conditioning on $\mathbf{m}$. Write $\mathbf{x} = \mathbf{m} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} := (\mathbf{I} - P)\mathbf{x}$. The pair $(\mathbf{m}, \boldsymbol{\epsilon})$ is a linear image of $\mathbf{x}$, hence jointly Gaussian, and $\text{Cov}(\mathbf{m}, \boldsymbol{\epsilon}) = P(\mathbf{I} - P)^\top = P - P^2 = 0$; jointly Gaussian and uncorrelated vectors are independent, so $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, (\mathbf{I} - P)(\mathbf{I} - P)^\top) = \mathcal{N}(\mathbf{0}, \mathbf{I} - P)$ is independent of $\mathbf{m}$, and $\mathbf{m} \sim \mathcal{N}(\mathbf{0}, P)$. In particular $m_j \sim \mathcal{N}(0, \pi_j)$ and $\epsilon_j \sim \mathcal{N}(0, 1 - \pi_j)$. For $j \notin \mathcal{K}$, $1 - \pi_j > 1 - \theta \geq 0$, and by independence

$$\mathbb{P}(x_j \geq \tau \mid \mathbf{m}) = \mathbb{P}(\epsilon_j \geq \tau - m_j \mid \mathbf{m}) = Q\left(\frac{\tau - m_j}{\sqrt{1 - \pi_j}}\right).$$

Since $\hat{i} = h(\mathbf{m})$ is $\mathbf{m}$ -measurable,

$$\mathbb{P}(x_i \geq \tau, \hat{i} \notin \mathcal{K} \mid \mathbf{m}) = \mathbb{1}\{h(\mathbf{m}) \notin \mathcal{K}\} \mathbb{P}(x_{h(\mathbf{m})} \geq \tau \mid \mathbf{m}) \leq \max_{j \notin \mathcal{K}} Q\left(\frac{\tau - m_j}{\sqrt{1 - \pi_j}}\right)$$

(with $\max \emptyset := 0$). On $F := \{\max_{j \notin \mathcal{K}} m_j \leq c\}$, for $j \notin \mathcal{K}$, $\tau - m_j \geq \tau - c \geq 0$ and $0 < \sqrt{1 - \pi_j} \leq 1$, so $(\tau - m_j)/\sqrt{1 - \pi_j} \geq \tau - m_j \geq \tau - c$, and since Q is decreasing the right-hand side is at most $Q(\tau - c)$. Taking expectations and bounding the conditional probability by 1 on F^c ,

$$\mathbb{P}(x_i \geq \tau, \hat{i} \notin \mathcal{K}) \leq Q(\tau - c) + \mathbb{P}(F^c).$$

For $j \notin \mathcal{K}$: if $\pi_j = 0$ then $m_j = 0$ a.s. and $\mathbb{P}(m_j > c) = 0$ (as $c \geq 0$); if $0 < \pi_j < \theta$ then $\mathbb{P}(m_j > c) = Q(c/\sqrt{\pi_j}) \leq Q(c/\sqrt{\theta})$, because $c/\sqrt{\pi_j} \geq c/\sqrt{\theta}$. The union bound over at most n indices gives $\mathbb{P}(F^c) \leq nQ(c/\sqrt{\theta})$. Inserting the three bounds into [Eq. \(103\)](#),

$$\mathbb{P}(\hat{i} = J) \leq \frac{r}{\theta n} + Q(\tau - c) + nQ(c/\sqrt{\theta}) + e^{-nQ(\tau)}$$

for every admissible (θ, c, τ) ; taking the infimum gives [Eq. \(36\)](#).

Distance. Let $\mathbf{y} := f(W\mathbf{x})$. If $\hat{i} \neq J$, then $\|\mathbf{y} - \mathbf{e}_{\hat{i}}\| \leq \|\mathbf{y} - \mathbf{e}_J\|$ because $\hat{i}$ is the vertex nearest to $\mathbf{y}$, and the triangle inequality gives $\sqrt{2} = \|\mathbf{e}_J - \mathbf{e}_{\hat{i}}\| \leq \|\mathbf{y} - \mathbf{e}_J\| + \|\mathbf{y} - \mathbf{e}_{\hat{i}}\| \leq 2\|\mathbf{y} - \mathbf{e}_J\|$. Hence $\|\mathbf{y} - \mathbf{T}(\mathbf{x})\| \geq \mathbb{1}\{\hat{i} \neq J\}/\sqrt{2}$ almost surely, and taking expectations, $\mathbb{E}\|f(W\mathbf{x}) - \mathbf{T}(\mathbf{x})\| \geq \mathbb{P}(\hat{i} \neq J)/\sqrt{2} \geq (1 - p_n(r))/\sqrt{2}$.

Asymptotics. Let $\tilde{s} = \sqrt{2 \log n}$ and take

$$\theta = \tilde{s}^{-2} = \frac{1}{2 \log n}, \quad c = 2, \quad \tau = \tilde{s} - w, \quad w = \frac{2 \log \tilde{s}}{\tilde{s}}.$$

For $n \geq 2$, $2 \log n > 1$, so $\theta \in (0, 1]$; as $w \rightarrow 0$ and $\tilde{s} \rightarrow \infty$, $\tau - c = \tilde{s} - w - 2 \geq 0$ for n large, so the choice is admissible. None of the choices depends on r . Term by term:

- $\frac{r}{\theta n} = \frac{2r \log n}{n}$.
- $c/\sqrt{\theta} = 2\tilde{s}$, and [Lemma C.3](#) gives $nQ(2\tilde{s}) \leq ne^{-2\tilde{s}^2} = ne^{-4 \log n} = n^{-3}$.
- With $\tau - c = \tilde{s} - (w + 2) \geq 0$, [Lemma C.3](#) and $(\tilde{s} - (w + 2))^2 \geq \tilde{s}^2 - 2\tilde{s}(w + 2)$ give

$$Q(\tilde{s} - w - 2) \leq e^{-\tilde{s}^2/2 + \tilde{s}(w+2)} = n^{-1} e^{2 \log \tilde{s} + 2\tilde{s}} = n^{-1} \tilde{s}^2 e^{2\tilde{s}},$$

whose logarithm $-\tilde{s}^2/2 + 2 \log \tilde{s} + 2\tilde{s} \rightarrow -\infty$.

- With $x := \tilde{s} - w > 0$, [Lemma C.6](#) and $n\varphi(\tilde{s} - w) = \frac{n}{\sqrt{2\pi}} e^{-\tilde{s}^2/2} e^{\tilde{s}w - w^2/2} = \frac{\tilde{s}^2 e^{-w^2/2}}{\sqrt{2\pi}}$ (as $ne^{-\tilde{s}^2/2} = 1$ and $\tilde{s}w = 2\log \tilde{s}$) give $nQ(\tau) \geq \frac{x}{1+x^2} \frac{\tilde{s}^2 e^{-w^2/2}}{\sqrt{2\pi}}$. For n large, $w \leq 1$ and $x \geq \tilde{s}/2 \geq 1$, so $\frac{x}{1+x^2} \geq \frac{1}{2x} \geq \frac{1}{2\tilde{s}}$ and $nQ(\tau) \geq \frac{\tilde{s} e^{-1/2}}{2\sqrt{2\pi}} \rightarrow \infty$; hence $e^{-nQ(\tau)} \rightarrow 0$.

Therefore $p_n(r) \leq \frac{2r \log n}{n} + \delta_n$ with $\delta_n := n^{-3} + n^{-1} \tilde{s}^2 e^{2\tilde{s}} + e^{-nQ(\tau)} \rightarrow 0$ independent of r . $\square$

F.8 Proof of [Corollary 6.9](#)

Proof. Idea. The floor chains [Theorem 5.2](#)(i)–(ii) with [Theorem 6.8](#): the error $\mathbb{E}\|\mathbf{T}_\theta - \mathbf{T}\|$ is at most $\varepsilon_{\text{bnd}} + R$, and, because $\mathbf{T}_\theta$ sees only $W\mathbf{x}_0$, at least $(1 - p_n(r))/\sqrt{2}$. The one gap is that f is not assumed measurable, and it is closed by replacing f with the continuous map $\mathbf{T}_\theta \circ W^+$, which agrees with f on the range of W . For the second statement it suffices that each $\mathbf{x}_0^\ell$ is a linear function of $W\mathbf{x}_0^\ell$, which is what injectivity gives.

The floor. Let $\mathbf{T}_\theta = f \circ W$ satisfy the hypotheses of [Theorem 5.2](#). Put $\tilde{f} := \mathbf{T}_\theta \circ W^+$, which is continuous, hence measurable; since $WW^+W = W$, $\tilde{f} \circ W = f \circ W \circ W^+ \circ W = f \circ W = \mathbf{T}_\theta$, so we may assume f measurable. By [Theorem 5.2](#)(i)–(ii), $\mathbb{E}_{\mu_0}\|\mathbf{T}_\theta - \mathbf{T}\| \leq \varepsilon_{\text{bnd}}(\theta) + R(\theta)$, and by [Theorem 6.8](#) the left-hand side $\mathbb{E}\|f(W\mathbf{x}_0) - \mathbf{T}(\mathbf{x}_0)\|$ is at least $(1 - p_n(r))/\sqrt{2}$.

The coarse-to-fine state. For $\mu_1 = \bigotimes \text{Unif}([v_k])$, [Theorem 4.9](#) gives $\mathbf{T}(\mathbf{x}_0)^{\ell,k} = e_{\arg \max_i x_{0,i}^{\ell,k}}$ for μ_0 -almost every $\mathbf{x}_0$, a function of $\mathbf{x}_0$, and for each position ℓ the blocks (ℓ, k) , $k \in [K]$, depend only on $\mathbf{x}_0^\ell \in \mathbb{R}^{\sum_k v_k}$. If W is injective on $\mathbb{R}^{\sum_k v_k}$, then $W^\top W$ is invertible, $W^+ = (W^\top W)^{-1} W^\top$ and $W^+W = \mathbf{I}$, so $\mathbf{x}_0^\ell = W^+W\mathbf{x}_0^\ell$ and $\mathbf{T} = \mathbf{T} \circ W^+ \circ W$ position by position: the transport map is a function of the read state $W\mathbf{x}_0$. $\square$

F.9 How tight is [Theorem 6.8](#)?

The bound $p_n(r)$ carries a factor $\log n$ that the truth does not. [Table 5](#) estimates the best recovery probability for two choices of W . Reading r coordinates exactly achieves $\approx r/n$ (the argmax is among the read coordinates with probability r/n , and it is then known); a random rank- r layer does worse, close to the random-energy value $n^{-(1-\sqrt{r/n})^2}$. We conjecture that r/n is optimal up to lower-order terms.

Table 5 Probability that the vertex nearest to the Bayes-optimal $f(W\mathbf{x}_0)$ equals the flat transport map $\arg \max_i x_{0,i}$ (Monte Carlo), against the bound $p_n(r)$ of [Theorem 6.8](#).

n	r	read r coordinates	random layer	r/n	$p_n(r)$
256	16	0.065	0.027	0.063	0.657
1,024	32	0.030	0.011	0.031	0.390
4,096	64	0.016	0.007	0.016	0.222
50,257	768	–	–	0.015	0.157

G Experimental Details

Threshold Monte Carlo ([Fig. 3](#)). For $n \in \{10, 10^2, 10^4, 10^6\}$ and 41 values of $\rho/s_n \in [0, 2]$, we draw $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and evaluate [Eq. \(11\)](#) with a stabilised log-sum-exp; 400 draws per point (120 for $n = 10^6$).

Commitment entropy ([Fig. 1, bottom](#)). For $v \in \{4, 8, 16, 24, 32\}$ and $V = v^3$, the flat token uses $a = 1$ and the three levels use $a = (3, 1.6, 1)$, all with $\sigma = 1$ and unit vertices (no energy matching). At 25 values of $t \in [0.02, 0.98]$ we sample $\mathbf{x}_0, \mathbf{x}_1$, compute the exact softmax posterior [Eq. \(10\)](#) and its entropy normalised by $\log V$ (flat) or $\log v$ (levels), averaging over 400 draws (240 for $V > 5000$) and 600 draws per level. Dotted lines are [Eq. \(15\)](#).

Exact fields ([Figs. 1, 4 and 5](#)). The field [Eq. \(5\)](#) is evaluated with uniform μ_1 , $\sigma = 0.55$ (0.6 for [Fig. 5](#)), and the time integral discretised on 200–400 points of $[0, 0.995]$ with log-sum-exp weights. Streamlines are integrated with a midpoint rule in the ambient space and projected onto the affine hull after each step; integration stops within 0.02 of a vertex. For $\Delta^2 \times \Delta^2$ the four-dimensional hull is drawn through the linear

map $(p, q) \mapsto (p^\top \Theta, 0) + 0.55 (q^\top \Theta) [\mathbf{u}_1 \ \mathbf{u}_2]^\top$, where $\Theta \in \mathbb{R}^{3 \times 2}$ holds the vertices of an equilateral triangle and $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^3$ are fixed tilted directions. The dynamics of Fig. 4 use 900 prior samples, explicit Euler steps of size 0.02 and absorption within 0.025 of a vertex.

Refinement (Fig. 6). Algorithm 1 is run with the exact per-level posterior Eq. (10) at probe time t_r in place of T_θ : at each round every uncommitted block draws fresh prior noise, forms its interpolant at $t_r \in \{0.18, 0.38, 0.58, 0.78\}$, and reads the posterior. Confidence is the largest posterior probability. Figure 6 uses one noise seed of the sentence “the map sends each point to the word”.

Cost of a training step (Table 4). One step consists of the transport loss Eq. (25) with the Jacobian–vector product computed by forward-mode differentiation with frozen parameters, the boundary loss at the data vertices, the anchor loss, the backward pass and an AdamW update. The network is the time-independent transformer map of all experiments (pre-norm blocks, feed-forward width $2d$), identical across codes except for the input layer $\mathbb{R}^D \rightarrow \mathbb{R}^d$ and the output layer $\mathbb{R}^d \rightarrow \mathbb{R}^D$; flat DBTM uses $D = V$, the codes $K = 2$ and $K = 3$ use $v_k = \lceil V^{1/K} \rceil$. Inputs are random interpolants (the cost does not depend on the data), sequence length 128. Each configuration runs in its own process on an Apple M5 GPU: 3 warm-up steps, then the median of 8 timed steps; memory is the peak driver allocation.

Trained models (Tables 2 and 3). The language has $C = W = 32$ clusters and words; for every cluster c the successor clusters are $S = 2$ clusters drawn without replacement with Dirichlet(2) weights, and likewise for every pair (cluster, previous word) the successor words; the first token is uniform. Training data are 200,000 sequences of length 8. The map is a pre-norm transformer with 3 layers, 4 heads, feed-forward width $2d$, a linear input layer on the state of each position, learned position embeddings and one softmax per block. Every step uses a batch of 256 sequences, AdamW with peak learning rate 10^{-3} and a one-cycle schedule (5% warm-up), gradient clipping at 1, and the loss $\mathcal{L}_{\text{transport}} + \mathcal{L}_{\text{bnd}} + \mathcal{L}_{\text{anchor}}$ with anchors at the thresholds t_k^* of Eq. (15) (computed with the vertex scale of the code). In half of the sequences each position has a uniformly random number of its leading levels pinned to the data (for the flat code, each position is pinned with probability $1/2$), as in the partial-context interpolant of DBTM. Refinement-in-loop training adds, with weight 1, the cross-entropy of Section 5 on 64 of the 256 sequences, rolled out with the token-level rule over $k = 4$ rounds. Samples: 2,000 sequences per model and sampler, $\kappa = 0.9$. The exponents $a = (1.92, 1.09)$ come from the greedy packing of Proposition 6.5 with $z = 1.28$, $v = 32$ and prior scale $\sqrt{2}$, and $a^* = 1.50$ from Eq. (34) with $n = V = 1,024$ and $\sigma = 1$. The level diagnostics in Section 7.2 use the saved standard-training models (flat $a = 1$, width 64; coarse-to-fine $a = (1, 1)$, width 96): the validity of a transition is split into the cluster transition $P(c' | c) > 0$ and the word transition $P(u' | c', u) > 0$, and the gap-filling probe pins the true tokens of the even positions as committed vertices, draws prior noise at the odd positions and decodes one evaluation of the map.