



DARTMOUTH

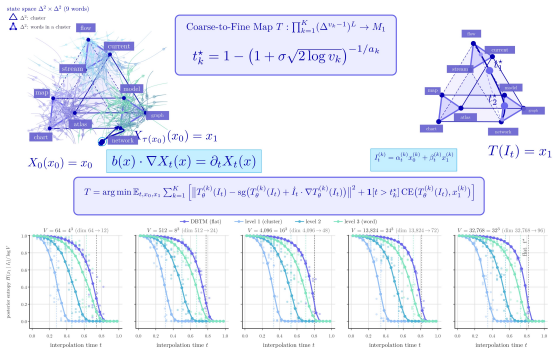
Department of Mathematics

Transport on Simplices

Coarse-to-fine maps for one-step language generation

Johannes Herter

Applied and Computational Mathematics
Seminar, Dartmouth College · 8 October 2026





1

Generative models for text

A language model is a distribution over sequences



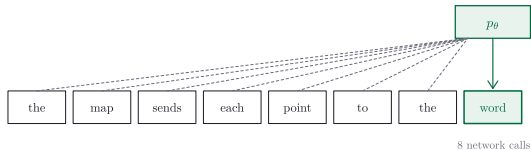
A sequence $x = (x^1, \dots, x^L)$ over a vocabulary of size $V \approx 5 \cdot 10^4$.

Autoregressive models factorise

$$p_{\theta}(x) = \prod_{\ell=1}^L p_{\theta}(x^{\ell} \mid x^1, \dots, x^{\ell-1})$$

and sample one token at a time.

The cost. L sequential network calls;
each waits for the previous one.

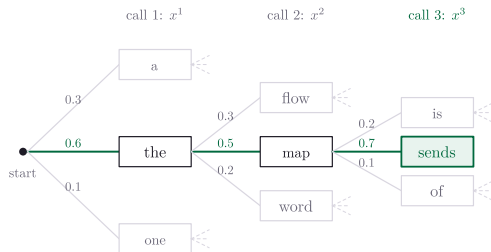


The chain rule and sequential sampling



- 1 **Conditional probability.** $p(A \cap B) = p(A)p(B | A)$ gives $p(x^1, \dots, x^L) = p(x^1)p(x^2, \dots, x^L | x^1)$.
- 2 **Induction on L .**
$$p(x) = \prod_{\ell=1}^L p(x^\ell | x^{<\ell}),$$

exact for every law on $\{1, \dots, V\}^L$. A network has to output L conditionals, each a distribution over V tokens, instead of one table with V^L entries.
- 3 **Sampling.** Draw $x^1 \sim p_\theta(\cdot)$, then $x^2 \sim p_\theta(\cdot | x^1)$, and so on. The input of call ℓ contains the output of call $\ell - 1$, so the L calls run one after another.
- 4 **Training.** On a data sequence every prefix $x^{<\ell}$ is given, and one pass with a causal mask evaluates all L terms of $\log p_\theta(x) = \sum_\ell \log p_\theta(x^\ell | x^{<\ell})$. Only sampling is sequential.



3 network calls, each after the previous one

Sampling walks down the tree of conditionals, one level per network call.

Masked diffusion models



step 4 of 4



each step: predict every mask in parallel, keep some

Masked diffusion (MDLM, LLaDA)

Mask each token with probability t ; predict all masked tokens at once: $\mathcal{L} = \mathbb{E} \frac{1}{t} \sum_{\ell: x_t^\ell = [\text{M}]} -\log p_\theta(x^\ell \mid x_t)$.

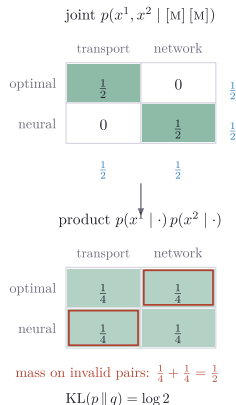
Limitation. Positions unmasked in one step are sampled *independently* (“the the”), so quality needs many steps.

Parallel unmasking and the independence error



- 1 **One call, all masks.** Given x_t , the network returns for every masked position ℓ the conditional law $p(x^\ell \mid x_t)$: one distribution per position, not their joint law.
- 2 **Several positions in one step.** Unmasking a set S of positions at once draws from the product $q(x^S) = \prod_{\ell \in S} p(x^\ell \mid x_t)$.
- 3 **The error.** With $p = p(x^S \mid x_t)$ and entropies H given x_t :

$$\text{KL}(p \parallel q) = \mathbb{E}_p \log p(x^S) - \sum_{\ell \in S} \mathbb{E}_p \log p(x^\ell) = \sum_{\ell \in S} H(x^\ell) - H(x^S),$$
 the total correlation; it is zero iff the positions in S are conditionally independent.
- 4 **Trade-off.** $|S| = 1$ is exact (the chain rule) and takes L steps; fewer steps mean larger sets S and a larger error.



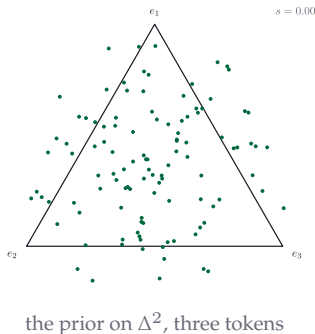
Continuous models on the simplex



Embed token i as the vertex e_i of the simplex $\Delta^{V-1} = \{p \in \mathbb{R}_{\geq 0}^V : \sum_i p_i = 1\}$. A sequence is a vertex of $(\Delta^{V-1})^L \subset \mathbb{R}^{VL}$; noise is $x_0 \sim \mathcal{N}(0, \sigma^2 I)$ on the whole ambient space.

The state moves *continuously* and all positions move *jointly*, through one vector field: a single step is possible in principle.

Simplex models. Dirichlet / Fisher flow, continuous diffusion: all transport a prior to the vertices, in many steps.

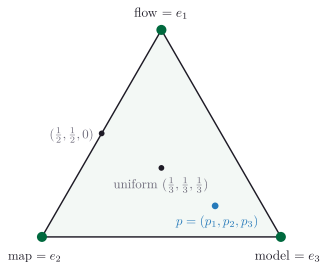


Points of the simplex are distributions over tokens



$\Delta^{V-1} = \{p \in \mathbb{R}_{\geq 0}^V : \sum_i p_i = 1\}$: the probability distributions on $\{1, \dots, V\}$.

- Vertex e_i : token i with probability one;
barycentre: the uniform law; edge $[e_i, e_j]$: the laws on $\{i, j\}$.
- $\|e_i - e_j\| = \sqrt{2}$ for all $i \neq j$: no token is closer to one token than to another.
- For a random vertex x_1 and any observation y , $\mathbb{E}[x_1 | y] = \sum_i \mathbb{P}(x_1 = e_i | y) e_i$ is the posterior over tokens, a point of Δ^{V-1} .
- A softmax output is a point of Δ^{V-1} , like these posteriors.



$$\Delta^2 = \{p \in \mathbb{R}_{\geq 0}^3 : p_1 + p_2 + p_3 = 1\}$$

$V = 3$ tokens: Δ^2

Flow matching



Interpolate $I_t = (1 - t)x_0 + tx_1$,
 $x_0 \sim \mu_0$, $x_1 \sim \mu_1$, and regress the
velocity:

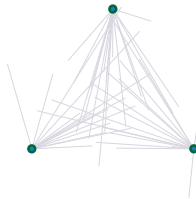
$$b_t(x) = \mathbb{E}[\dot{I}_t \mid I_t = x],$$

learned by $\min_{\theta} \mathbb{E} \|v_{\theta}(t, I_t) - \dot{I}_t\|^2$.

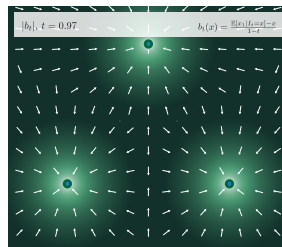
Transport

The ODE $\dot{X} = b_t(X)$, $X_0 \sim \mu_0$, has
 $X_t \sim \text{Law}(I_t)$ for all t ; in particular
 $X_1 \sim \mu_1$.

$t = 1.00$



$I_t = (1 - t)x_0 + tx_1$



Left: pairs (x_0, x_1) on straight lines. Right: the exact field b_t ,
three basins.

Why. $\frac{d}{dt} \mathbb{E} \phi(I_t) = \mathbb{E}[\nabla \phi \cdot \dot{I}_t] = \mathbb{E}[\nabla \phi \cdot b_t(I_t)]$, so
both laws solve $\partial_t \mu_t + \nabla \cdot (\mu_t b_t) = 0$.

Flow matching on the simplex



1 **Schedule.** $I_t = \alpha_t x_0 + (1 - \alpha_t)x_1$ with $\alpha_t = (1 - t)^a$ (the previous slides: $a = 1$), so $\dot{\alpha}_t = -\frac{a}{1-t} \alpha_t$.

2 **Eliminate the noise.** $x_0 - x_1 = (I_t - x_1)/\alpha_t$, hence

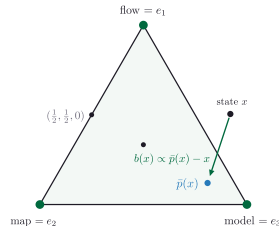
$$\dot{I}_t = \dot{\alpha}_t(x_0 - x_1) = \frac{a}{1-t} (x_1 - I_t).$$

3 **Condition on $I_t = x$.** $b_t(x) = \mathbb{E}[\dot{I}_t \mid I_t = x] = \frac{a}{1-t} (\mathbb{E}[x_1 \mid I_t = x] - x)$.

4 **The mean is a distribution.** x_1 is a vertex, so $\mathbb{E}[x_1 \mid I_t = x] = \sum_i \mathbb{P}(x_1 = e_i \mid I_t = x) e_i =: p_t(x) \in \Delta^{V-1}$, the posterior over tokens.

5 **Velocity.**

$$b_t(x) = \frac{a}{1-t} (p_t(x) - x) :$$
 every state moves straight towards the current posterior over tokens.



$$\Delta^2 = \{p \in \mathbb{R}_{\geq 0}^3 : p_1 + p_2 + p_3 = 1\}$$

Steps 1–3 hold for every data law; the only unknown is the posterior $p_t(x)$, learned as a softmax. Averaged over t the field points to one tilted mean $\bar{p}(x)$ (Lemma 3.6).

One-step generation and the end-point map



Sampling integrates $\dot{X} = b_t(X)$ with K Euler steps. The single pairs move straight, but the *average* field bends: trajectories are curved, and one Euler step lands in the middle of the simplex.

One-step generation needs the end point. $T(x_0) := X_1(x_0)$, the end of the trajectory. Distillation and consistency models learn T from a trained multi-step teacher.

Beckmann transport [Lee et al. 2026]; DBTM [Tang-Wang 2026]

Average the velocity over time as well: the field $b(x) = \mathbb{E}_t[\dot{I}_t \mid I_t = x]$ has no clock. Its end-point map is constant along trajectories, which gives an equation for T that can be trained *without a teacher*.

The rest of the talk: this construction, its theory on products of simplices, and what the theory changes in practice.



2

Beckmann transport

Optimal transport: couplings and fluxes

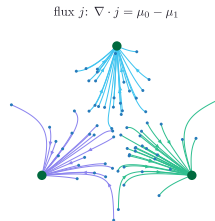
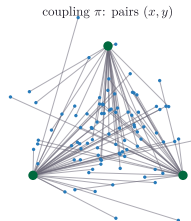


Kantorovich problem

$W_1(\mu_0, \mu_1) = \min_{\pi} \int |x - y| d\pi$ over couplings π of μ_0, μ_1 . Unknown: a measure on $\mathbb{R}^d \times \mathbb{R}^d$.

Beckmann's problem (1952)

$B(\mu_0, \mu_1) = \inf \{ \int |j| dx : \nabla \cdot j = \mu_0 - \mu_1 \}$. Unknown: a vector field j on $\mathbb{R}^d$.



Gaussian μ_0 (blue), three atoms μ_1 (green). Left: the independent coupling. Right: the flux $j = \int_0^1 \mu_t b_t dt$ of flow matching. It gives the direction and amount of mass through each point, not which source point goes to which target point.

Full proof: Beckmann's problem equals W_1



1 **Weak form.** $\nabla \cdot j = \mu_0 - \mu_1$ means
$$\int \nabla \varphi \cdot j \, dx = \int \varphi \, d\mu_1 - \int \varphi \, d\mu_0 \quad \text{for all } \varphi \in C_c^1.$$

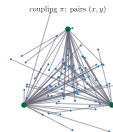
2 **A coupling gives a flux.** For $\pi \in \Pi(\mu_0, \mu_1)$ move mass along segments:
$$\int \psi \cdot j_\pi := \iint_0^1 \psi(x + s(y-x)) \cdot (y-x) \, ds \, d\pi(x, y).$$

3 **It is admissible.** With $\psi = \nabla \varphi$ the inner integral is $\varphi(y) - \varphi(x)$, so $\int \nabla \varphi \cdot j_\pi = \int \varphi \, d\mu_1 - \int \varphi \, d\mu_0$.

4 **Its mass.** $\int |j_\pi| \leq \iint_0^1 |y-x| \, ds \, d\pi = \int |x-y| \, d\pi$.
Minimising over π :
$$B(\mu_0, \mu_1) \leq W_1(\mu_0, \mu_1).$$

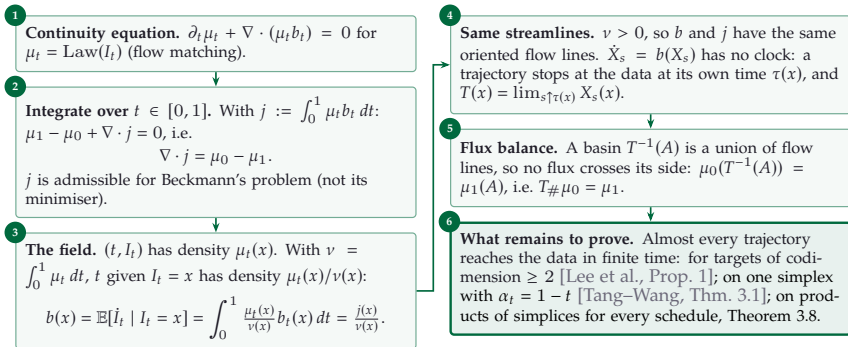
5 **Converse.** For admissible j and $\varphi \in C^1$ with $|\nabla \varphi| \leq 1$: $\int \varphi \, d(\mu_1 - \mu_0) = \int \nabla \varphi \cdot j \leq \int |j|$. Kantorovich-Rubinstein duality, $W_1 = \sup_{\text{Lip } \varphi \leq 1} \int \varphi \, d(\mu_1 - \mu_0)$ (smooth φ suffice, by mollification and cut-off), gives $W_1 \leq B$.

6 **Conclude.** $B(\mu_0, \mu_1) = W_1(\mu_0, \mu_1)$ [Santambrogio 2015, Ch. 4]. The flux lives on $\mathbb{R}^d$, the coupling on $\mathbb{R}^d \times \mathbb{R}^d$. ■



a coupling moves mass
along segments

Time-averaged flow matching gives a Beckmann flux



flux j : $\nabla \cdot j = \mu_0 - \mu_1$



streamlines of j , coloured

by end point

DBTM: interpolant and autonomous field



A token is a vertex e_i of $\Delta^{V-1} \subset \mathbb{R}^V$, a sentence a vertex of $(\Delta^{V-1})^L$. Data $\mathbf{x}_1 \sim \mu_1$, noise $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma^2 I)$.

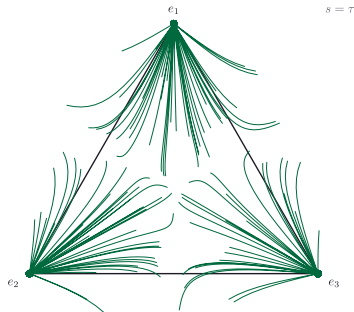
Interpolant

$$\mathbf{I}_t = \alpha_t \mathbf{x}_0 + (1 - \alpha_t) \mathbf{x}_1, \quad \alpha_t = (1 - t)^a, \quad t \sim \mathcal{U}[0, 1].$$

Autonomous field

$$\mathbf{b}(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{I}}_t \mid \mathbf{I}_t = \mathbf{x}], \text{ averaged over } t \text{ as well.}$$

The flow $\dot{\mathbf{X}}_s = \mathbf{b}(\mathbf{X}_s)$ has no clock; a sample is where it stops, $\mathbf{T}(\mathbf{x}_0) = \lim_{s \uparrow \tau} \mathbf{X}_s$.



The autonomous flow on Δ^3



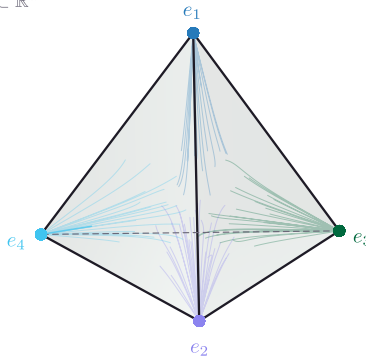
Four tokens: $\Delta^3 \subset \mathbb{R}^4$ is a tetrahedron, and $\mathbf{b}$ is tangent to its affine hull.

Particles started near the barycentre are moved by the *exact* field ($a = 1$, $\sigma = 0.6$); faint lines are complete trajectories of a second sample.

Each particle runs into one vertex and stops there, at a finite flow time τ (Theorem 3.8). The colour marks the vertex it reaches: the four basins of $\mathcal{T}$.

$\Delta^3 \subset \mathbb{R}^4$

$s = \tau$



The end-point map: what is known and what is not



Known: the field

$b(x) = \mathbb{E}[\dot{I}_t \mid I_t = x]$: a regression target, learned from samples of the interpolant without solving an ODE.

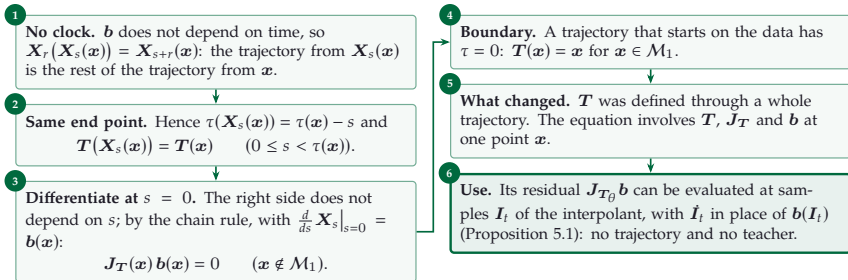
Why it is hard. $T(x)$ depends on the whole trajectory, up to its own time $\tau(x)$. Regressing T needs pairs $(x, T(x))$ from a teacher; the interpolant's pairs are independent, $x_1 \neq T(x_0)$.

Unknown: the end point

$T(x) = \lim_{s \uparrow \tau(x)} X_s(x)$: where the trajectory from x meets the data.

The way around it. T is constant along trajectories. This gives an equation for T at single points (next slide), and single points I_t are what the interpolant provides.

Derivation of the conservation equation



trajectories of $\mathbf{b}$ before
they reach a vertex

The DBTM training loss



The end-point map is the solution of a first-order equation along the field:

Conservation equation (Tang–Wang, Prop. 3.2)

$$b(x) \cdot \nabla T(x) = 0 \quad (x \notin \mathcal{M}_1), \quad T(x) = x \quad (x \in \mathcal{M}_1).$$

The residual is evaluated on samples of the interpolant, without integrating the flow:

$$\mathcal{L}_{\text{transport}} = \mathbb{E} \left\| T_{\theta}(I_t) - \text{sg}(T_{\theta}(I_t) + \nabla T_{\theta}(I_t) \dot{I}_t) \right\|^2,$$

plus a boundary loss on the data and an **anchor** $\text{CE}(T_{\theta}(I_t), x_1)$ switched on after the token is decided, at a time $t^{\star}$ that DBTM derives from the random-energy model as $V \rightarrow \infty$.

Open questions for flat DBTM



Convergence. Proven for equal simplices and the linear schedule only. Does the flow reach the data for other schedules?

Schedule. $a = 1$ is a choice. How much training does it spend on the decision that matters?

Commitment. $t^\star$ comes from a heuristic in the limit $V \rightarrow \infty$. When is a token decided at finite V ?

State. Every position carries $V \approx 5 \cdot 10^4$ coordinates, read through a layer of width 768.

All four questions have the same setting: Beckmann transport on a *product* of simplices.



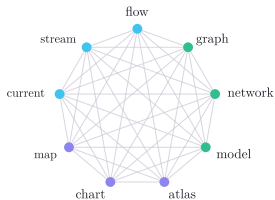
3

Products of simplices

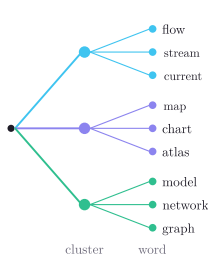
Coarse-to-fine: the cluster first, then the word



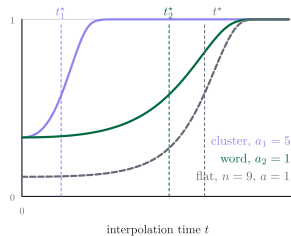
flat: Δ^8 , all $\|e_i - e_j\| = \sqrt{2}$



code: $(\text{cluster}, \text{word}) \in \Delta^2 \times \Delta^2$



$\mathbb{E} \pi_\star$ (mass on the true symbol)



Flat. One simplex Δ^{V-1} , every pair of tokens at distance $\sqrt{2}$. The posterior over all V tokens resolves in one transition, at $t^\star$ (Corollary 4.4).

Code. $w \mapsto (c^{(1)}(w), c^{(2)}(w))$, a vertex of $\Delta^2 \times \Delta^2$. The chain rule over levels, $p(w) = p(c^{(1)})p(c^{(2)} | c^{(1)})$, in place of the chain rule over positions.

Schedules. With $\alpha_k(t) = (1 - t)^{a_k}$ and $a_1 > a_2$ the noise of level 1 vanishes first:
 $t_1^\star = 0.14 < t_2^\star = 0.54$; flat:
 $t^\star = 0.67$ (uniform data, $\sigma_k = 1$).

Hierarchical codes and per-level schedules



Definition 3.1 (hierarchical code)

An injective map $\gamma : \{1, \dots, V\} \rightarrow \prod_{k=1}^K [v_k]$, $w \mapsto (c^{(1)}(w), \dots, c^{(K)}(w))$, $\prod_k v_k \geq V$; level 1 is coarsest. Token w at position ℓ is the stacked vertex $(e_{c^{(1)}(w)}, \dots, e_{c^{(K)}(w)}) \in \prod_k \Delta^{v_k-1}$.

Assumption 3.2 (coarse-to-fine interpolant)

Level k has a schedule $\alpha_k(t) = (1-t)^{a_k}$, $a_k > 0$, and a prior scale σ_k . With $t \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0 \sim \mathcal{N}(0, \Sigma)$, $\mathbf{x}_1 \sim \mu_1$ independent:

$$\mathbf{I}_t^{(\ell, k)} = \alpha_k(t) \mathbf{x}_0^{(\ell, k)} + (1 - \alpha_k(t)) \mathbf{x}_1^{(\ell, k)}.$$

Flat DBTM is $K = 1$.

Proposition 3.5 (ambient dimension)

If $\prod_k v_k = V$, the state has $L \sum_k v_k \geq LKV^{1/K}$ coordinates, equality iff $v_k = V^{1/K}$; over real K , $KV^{1/K}$ is smallest at $K = \ln V$, with value $e \ln V$.

GPT-2 ($V = 50,257$): $e \ln V \approx 29.4$; the code $K = 3$, $v_k = 37$ uses 111 coordinates per position instead of 50,257.

Proposition 3.3: the exact field



Given t and the end point e_j , the interpolant is Gaussian:

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

$$w_j(t, \mathbf{x}) = \frac{\pi_j p(\mathbf{x} \mid t, j)}{q(\mathbf{x})},$$

with $q(\mathbf{x}) = \int_0^1 \sum_j \pi_j p(\mathbf{x} \mid s, j) ds$; w_j is the posterior of (time, end point) given $\mathbf{I}_t = \mathbf{x}$.

Proposition 3.3 (exact field)

For every $\mathbf{x} \notin \mathcal{M}_1$ the integrals converge, $\mathbf{b} \in C^\infty(\mathbb{R}^D \setminus \mathcal{M}_1)$, and block (ℓ, k) of the field is

$$\mathbf{b}^{(\ell, k)}(\mathbf{x}) = \int_0^1 \sum_j w_j(t, \mathbf{x}) \frac{a_k}{1-t} (\mathbf{e}_{j_{\ell, k}} - \mathbf{x}^{(\ell, k)}) dt.$$

Idea. Given (t, j) the velocity is $\dot{\mathbf{I}}_t^{(\ell, k)} = \frac{a_k}{1-t} (\mathbf{e}_{j_{\ell, k}} - \mathbf{I}_t^{(\ell, k)})$ (differentiate α_k and eliminate $\mathbf{x}_0$); the conditional expectation is this velocity averaged against the posterior w_j .

Lemma 3.6 and Proposition 3.4: structure of the field



Lemma 3.6 (common speed)

With $M(\mathbf{x}) = \int_0^1 \sum_j \frac{w_j}{1-t} dt \geq 1$ and the tilted posterior mean $\bar{\mathbf{p}}^{(\ell,k)}(\mathbf{x}) \in \Delta^{v_k-1}$,

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

Every block relaxes towards a point of its own simplex at rate $a_k M$; one scalar speed M for all blocks.

Proposition 3.4 (structure of the field)

- (i) *Convex pull*: $\mathbf{b}^{(\ell,k)} = \sum_i c_i^{(\ell,k)} (\mathbf{e}_i - \mathbf{x}^{(\ell,k)})$ with $c_i^{(\ell,k)} \geq 0$.
- (ii) *Hull attraction*: $h = \mathbf{1}^\top \mathbf{X}_s^{(\ell,k)} - 1$ obeys $\dot{h} = -C^{(\ell,k)} h$; the affine hull is invariant.
- (iii) *Forward invariance*: on a face $x_i^{(\ell,k)} = 0$, $b_i^{(\ell,k)} = c_i^{(\ell,k)} \geq 0$; the product of simplices is never left.

Theorem 3.7: absorption near a vertex



Measure the distance to a vertex $e_{j^\star}$ level by level,
 $V(\mathbf{x}) = \max_{\ell,k} \|\mathbf{x}^{(\ell,k)} - e_{j_{\ell,k}^\star}\|^{1/a_k}$.

Theorem 3.7 (local absorption)

There are $\varepsilon \in (0, 1)$ and $0 < c \leq C$ such that for $0 < V(\mathbf{x}_0) < \varepsilon$: the flow stays in $\{V < \varepsilon\}$, $V(\mathbf{X}_s)$ decreases strictly, and $\mathbf{X}_s \rightarrow e_{j^\star}$ at a finite time τ with

$$\frac{V(\mathbf{x}_0)}{C} \leq \tau(\mathbf{x}_0) \leq \frac{V(\mathbf{x}_0)}{c},$$
$$\|\mathbf{X}_s^{(\ell,k)} - e_{j_{\ell,k}^\star}\| \leq (C(\tau - s))^{a_k}.$$

Idea. Near the vertex the time average produces a speed $M \asymp 1/V$ that blows up, so $\frac{d}{ds}V \leq -c$: absorption in finite time even though $\dot{\alpha}_k(t) \rightarrow 0$ for $a_k > 1$.

In the last stretch a level with a larger exponent is polynomially closer to its vertex.

Theorem 3.8: convergence for every schedule



Theorem 3.8 (convergence and transport)

Under Assumption 3.2, for any $a_k > 0$, $\sigma_k > 0$ and $v_k \geq 2$:

- (i) for μ_0 -a.e. x the flow reaches $\mathcal{M}_1^+$ at a finite time $\tau(x)$, and $\mathbb{E}_{\mu_0} \tau \leq 1$;
- (ii) $\mathcal{A} = \{\tau < \infty\}$ is open of full measure, $T = \lim_{s \uparrow \tau} X_s$ is locally constant on $\mathcal{A}$, so $T \in C^\infty(\mathcal{A})$ and $J_T b = 0$ there;
- (iii) $T_{\#} \mu_0 = \mu_1$;
- (iv) (occupation) $\mathbb{E}_{\mu_0} \int_0^\tau \phi(X_s) ds \leq \int \phi q$ for every Borel $\phi \geq 0$.

Local part. Theorem 3.7: trajectories that come close to a data point are absorbed.

Global part. The current qb has $\nabla \cdot (qb) = \mu_0 - \mu_1$; Smirnov's theorem decomposes it into curves from μ_0 to μ_1 , whose total time is at most $\int q = 1$.

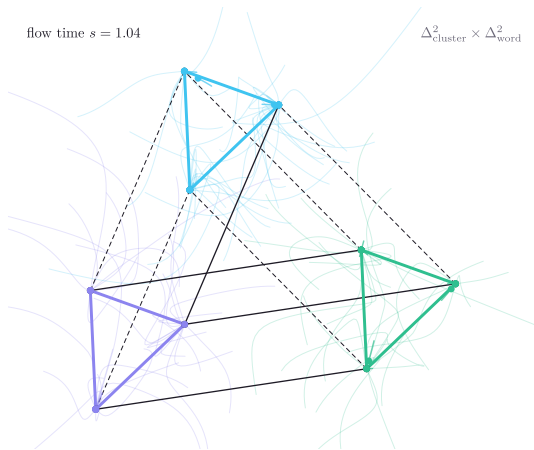
Numerically $\mathbb{E} \tau = 1.000 \pm 0.014$ on Δ^2 : the bound is attained.

The exact flow on $\Delta^2 \times \Delta^2$



Nine words, coded as (cluster, word) with $a = (5, 1)$. Sixty prior particles moved by the *exact* field of Proposition 3.3.

Every particle first enters the word triangle of a cluster (coloured), and only then resolves a word: the flow is coarse-to-fine. All arrive by $s \approx 1$, consistent with $\mathbb{E}\tau \leq 1$.



The boundary of $\Delta^2 \times \Delta^2$: two solid tori

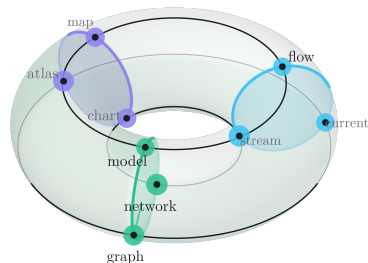


$\partial(\Delta^2 \times \Delta^2) = (\partial\Delta^2 \times \Delta^2) \cup (\Delta^2 \times \partial\Delta^2)$
is two solid tori glued along the torus
 $\partial\Delta^2 \times \partial\Delta^2$: a genus-one splitting of
 S^3 . The torus carries the nine words
and the 3×3 squares edge $\times$ edge.

A cluster face $\{e_a\} \times \Delta^2_{\text{word}}$ is a
meridian disk of the inner solid torus.
Under the exact flow ($a = (5, 1)$) a
particle first reaches the disk of its
cluster, then moves inside it to a word;
the discs at the words count arrivals.

flow time $s = 1.47$

clusters 70/70, words 70/70

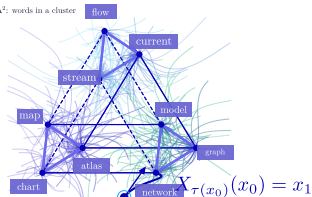


Cluster disks drawn through a radial homeomorphism $\Delta^2 \rightarrow D^2$.

state space $\Delta^2 \times \Delta^2$ (9 words)

$\triangle$: Δ^2 : cluster

$\triangle$: Δ^2 : words in a cluster



$$X_{\tau(x_0)}(x_0) = x_1$$

$$X_0(x_0) = x_0$$

$$b(x) \cdot \nabla X_t(x) = \partial_t X_t(x)$$

FIGURE 1, STEP BY STEP

1 The exact field on $\Delta^2 \times \Delta^2$ (9 words, 3 clusters) and its transport equation $b \cdot \nabla X_t = \partial_t X_t$; a path $x_0 \mapsto x_1$.

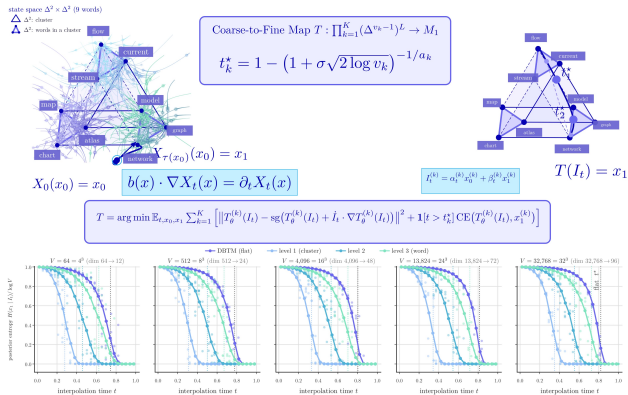
2 The coarse-to-fine map T and the commitment times $t_k^\star$.

3 The interpolant $I_t^{(k)}$ towards *network*: $T(I_t) = x_1$.

4 The training objective for T_θ .

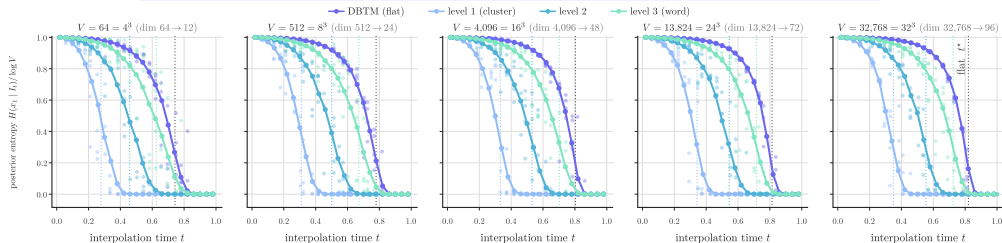
5 Posterior entropy: clusters are decided first, then words.

Figure 1



The exact field moves particles to a cluster, then a word; the interpolant commits at t_1^* , then t_2^* .

Posterior entropy: flat DBTM and a three-level code



Monte Carlo $H(x_1 | I_t) / \log V$ for flat DBTM (one schedule) and the three levels of a code, at $V = 4^3, \dots, 32^3$.

Dotted: the commitment times $t_k^\star$ of Theorem 4.5. Clusters are decided first, then words.



4

Commitment times

Proposition 4.1: the posterior factorises



Proposition 4.1 (factorisation)

Fix t . If $\mu_1 = \bigotimes_{\ell,k} \mu_1^{(\ell,k)}$, the posterior of the end point factorises over blocks. If each $\mu_1^{(\ell,k)}$ is uniform, every factor is a softmax with inverse temperature $\beta_k(t) = \frac{1-\alpha_k(t)}{\alpha_k(t)^2 \sigma_k^2}$:

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

Proof. The likelihood is a product over blocks, so is the prior; Bayes gives a product. In each factor $\|\mathbf{x} - (1-\alpha)\mathbf{e}_i\|^2 = \|\mathbf{x}\|^2 - 2(1-\alpha)x_i + (1-\alpha)^2$, and only $-2(1-\alpha)x_i$ depends on i .

Caveat. The factorisation holds at fixed t . The autonomous field mixes over t , and does not factorise.

Lemma C.1: the posterior of the true vertex



One block, n symbols, true symbol $i^\star$. Write $\mathbf{I}_t = (1 - \alpha)\mathbf{e}_{i^\star} + \alpha\sigma\mathbf{g}$ with $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$ and the signal-to-noise ratio $\rho = \frac{1-\alpha}{\alpha\sigma}$. Then

Lemma C.1 (log-odds representation)

$$\pi_\star = \mathbb{P}(\mathbf{x}_1 = \mathbf{e}_{i^\star} \mid \mathbf{I}_t) = \left(1 + \sum_{j \neq i^\star} e^{\rho(g_j - g_{i^\star}) - \rho^2}\right)^{-1}.$$

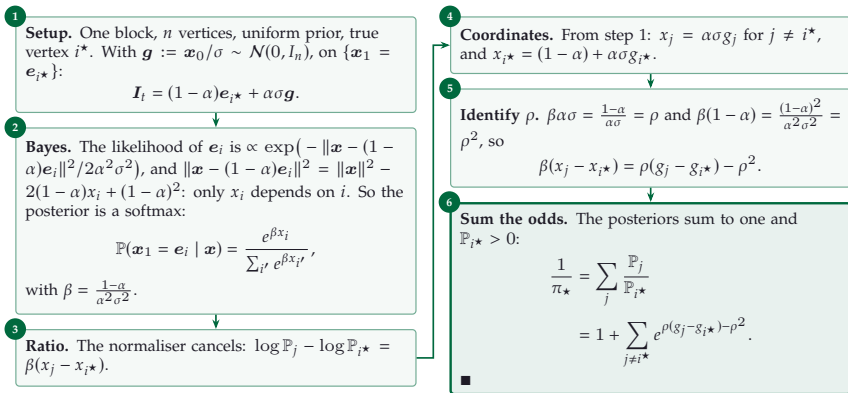
The true symbol has the advantage ρ^2 , competitor j the random score ρg_j . The posterior concentrates on the true symbol once ρ^2 exceeds the *largest* of the $n - 1$ scores, and

$$\max_{j \neq i^\star} g_j \approx s_n := \sqrt{2 \log(n-1)}.$$

So the transition sits at $\rho \approx s_n$.

Why $\log n$. A vocabulary 50× larger moves the threshold by only $\sqrt{2 \log 50} \approx 2.8$ in $s_n^2/2$.

Full proof of Lemma C.1



the posterior over n
symbols; the true one in
colour

Theorem 4.2: sharp threshold



Theorem 4.2 (sharp threshold)

For $u, u' > 0$:

above, $\rho \geq s_n + 2u$:

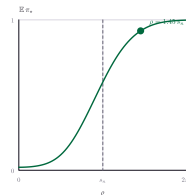
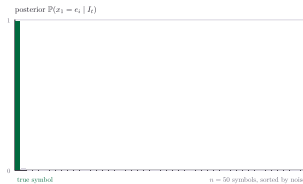
$$1 - \pi_\star \leq e^{u' - \rho(\rho - s_n - 2u)}$$

w.p. $\geq 1 - 2e^{-u^2/2} - e^{-u'}$;

below, $\rho \leq s_n - h - u - \eta$:

$$\pi_\star \leq e^{-\rho\eta}$$

w.p. $\geq 1 - e^{-u^2/2} - e^{-\Lambda_n(h)}$.



Corollary 4.4: the commitment time



Corollary 4.4 (commitment time)

For $\alpha(t) = (1-t)^a$, $\rho(t) = \frac{(1-t)^{-a}-1}{\sigma}$ is an 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

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

increasing in n and σ , decreasing in a . By Theorem 4.2 the interpolant has committed w.h.p. for $t \geq t(s_n + 2u + \delta)$ and not for $t \leq t(s_n - h - u - \eta)$.

Proof. Solve $\rho = ((1-t)^{-a} - 1)/\sigma$ for t ; ρ is increasing in t , so $t(\cdot)$ is increasing, and monotonicity in n, σ, a is read off the formula.

Relation to DBTM. This is DBTM's REM formula with $\log V$ replaced by $\log(n-1)$, now with a two-sided guarantee at every n .

Proof idea for Theorem 4.2



Everything is about $S = \sum_{j \neq i^*} e^{\rho g_j}$, since $\pi_\star = 1/(1 + e^{-\rho g_{i^*} - \rho^2} S)$.

Above: S is small. The mean of S is carried by values $g_j \approx \rho$ that no competitor reaches. Truncate at $s_n + u$, which the maximum exceeds with probability $\leq e^{-u^2/2}$, and apply Markov's inequality to the truncated sum:

$$\begin{aligned}\mathbb{E}S_M &= m e^{\rho^2/2} \Phi(M - \rho) \\ &\leq m e^{\rho M - M^2/2}.\end{aligned}$$

Below: S is large. One term suffices. By independence and the Mills ratio the largest competitor exceeds $s_n - h$ with probability $\geq 1 - e^{-\Lambda_n(h)}$, and then this single term already outweighs e^{ρ^2} .

The REM heuristic of DBTM becomes two bounds with explicit failure probabilities.

Full proof of Theorem 4.2: five Gaussian facts



1 **Upper tail.** $\mathbb{E}e^{\lambda g} = \int e^{\lambda z} \varphi(z) dz = e^{\lambda^2/2}$ (complete the square). Markov: $Q(x) \leq e^{-\lambda x} \mathbb{E}e^{\lambda g} = e^{-\lambda x + \lambda^2/2}$; take $\lambda = x$:

$$Q(x) \leq e^{-x^2/2} \quad (x \geq 0).$$

2 **Maximum, upper tail.** Union bound and step 1: $\mathbb{P}(\max_{i \leq m} g_i > s + u) \leq mQ(s + u) \leq me^{-(s+u)^2/2} \leq me^{-s^2/2} e^{-u^2/2} = e^{-u^2/2}$, since $(s + u)^2 \geq s^2 + u^2$ and $s^2/2 = \log m$.

3 **Truncated moment.** $\mathbb{E}[e^{\rho g} \mathbf{1}\{g \leq M\}] = \int_{-\infty}^M e^{\rho z} \varphi(z) dz = e^{\rho^2/2} \int_{-\infty}^M \varphi(z - \rho) dz = e^{\rho^2/2} \Phi(M - \rho)$.

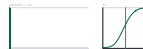
4 **Mills ratio.** $r(x) = Q(x) - \frac{x}{1+x^2} \varphi(x)$ has $r'(x) = -\frac{2\varphi(x)}{(1+x^2)^2} < 0$ and $r(\infty) = 0$, so $r > 0$:

$$Q(x) \geq \frac{x}{1+x^2} \varphi(x) \quad (x > 0).$$

5 **Maximum, lower tail.** By independence and $1 - y \leq e^{-y}$:

$$\mathbb{P}(\max_{i \leq m} g_i < x) = (1 - Q(x))^m \leq e^{-mQ(x)}.$$

6 **Use.** Steps 1–3 bound the competitors' sum S from *above* (commitment); steps 4–5 bound its largest term from *below* (no commitment). ■

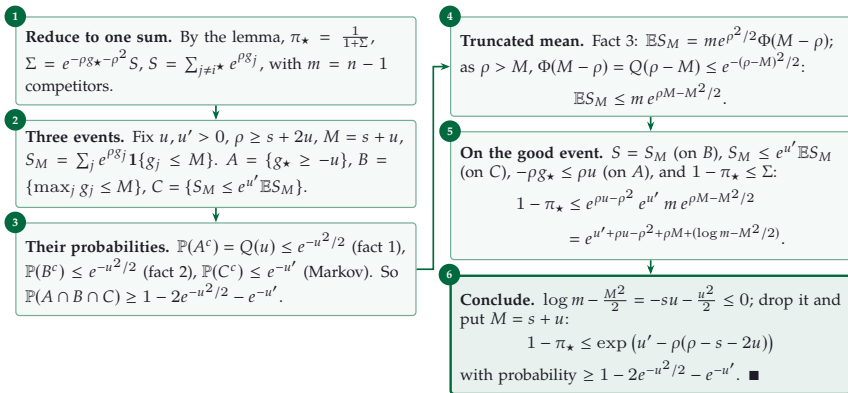


$$g \sim \mathcal{N}(0, 1),$$

$$Q(x) = \mathbb{P}(g > x),$$

$$s = \sqrt{2 \log m}$$

Full proof of Theorem 4.2: above the threshold



above $\rho = s_n$ the truth
takes the mass

Full proof of Theorem 4.2: below the threshold



1 **Setup.** Fix $u > 0$, $h \in (0, s)$, $\eta > 0$, $\rho \leq s - h - u - \eta$. Let $G = \max_{j \neq i^*} g_j$, $D = \{g_\star \leq u\}$, $E = \{G \geq s - h\}$.

2 $\mathbb{P}(D^c)$. $\mathbb{P}(g_\star > u) = Q(u) \leq e^{-u^2/2}$ (fact 1).

3 $\mathbb{P}(E^c)$. With $x = s - h > 0$, facts 5 and 4:
 $\mathbb{P}(G < x) \leq e^{-mQ(x)}$

$$\leq \exp\left(-m \frac{x}{1+x^2} \varphi(x)\right).$$

4 **The exponent is $\Lambda_n(h)$.** $m\varphi(x) = \frac{m}{\sqrt{2\pi}} e^{-s^2/2} e^{sh-h^2/2} = \frac{e^{sh-h^2/2}}{\sqrt{2\pi}}$, so $\mathbb{P}(D \cap E) \geq 1 - e^{-u^2/2} - e^{-\Lambda_n(h)}$.

5 **One term suffices.** With $z = \rho(G - g_\star) - \rho^2$, one term of Σ is e^z : $\pi_\star = \frac{1}{1+\Sigma} \leq \frac{1}{1+e^z} \leq e^{-z}$.

6 **Conclude.** On $D \cap E$: $G - g_\star \geq s - h - u$, so $z \geq \rho(s - h - u - \rho) \geq \rho\eta$ and
 $\pi_\star \leq e^{-\rho\eta}$
 with probability $\geq 1 - e^{-u^2/2} - e^{-\Lambda_n(h)}$. ■



below $\rho = s_n$ one
competitor wins

Theorem 6.1: the commitment profile



Theorem 6.1 (commitment profile)

Write $\rho = s_n + x$ and let $\Lambda = \log \frac{\pi_\star}{1-\pi_\star}$. For $y > 0$, $w \in (0, s_n)$, with probability $\geq 1 - e^{-y} - e^{-\Lambda_n(w)}$:

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

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

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

Hence $\sup_{|x| \leq X} |\mathbb{E} \pi_\star(s_n + x) - \Phi(x)| \rightarrow 0$ and $\Lambda/s_n - (x + g_\star) \rightarrow 0$ in probability.

Idea. $\Lambda = \rho g_\star + \rho^2 - \log S$; bound $\log S$ above by Markov and below by its largest term. What stays random is $g_\star$, a standard Gaussian: it produces Φ .

From 10% to 90% while ρ crosses $s_n \pm 1.28$, for every vocabulary. A single sample commits over a window of length $O(1/s_n)$ at $\rho = s_n - g_\star$.

The threshold sharpens as the vocabulary grows

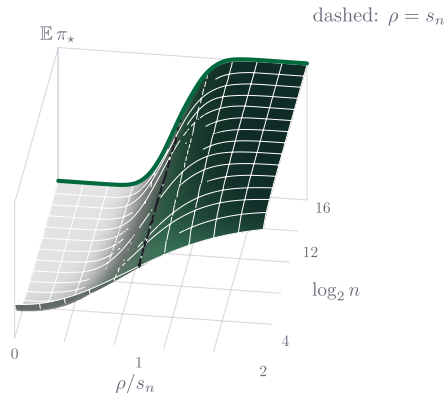


$\mathbb{E} \pi_\star$ as a function of ρ/s_n and $\log_2 n$, by Monte Carlo, for $n = 4, \dots, 2^{16}$.

In ρ the transition has width $O(1)$ (Theorem 6.1: $\mathbb{E} \pi_\star(s_n + x) \rightarrow \Phi(x)$), so in ρ/s_n it has width $O(1/s_n)$: the profile tends to a step at $\rho = s_n$.

Dashed: $\rho = s_n$. Green: the profile at one n .

$n = 65,536$



Theorem 4.5: levels commit in order



Theorem 4.5 (staggered commitment)

Let $\mu_1 = \bigotimes_{\ell,k} \text{Unif}([v_k])$. Each block's $\pi_{\star}^{(\ell,k)}(t)$ has the law of the lemma with $n = v_k$, $\rho = \rho_k(t)$, blocks independent; Theorem 4.2 holds per block with

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

all L positions of a level simultaneously w.p. $\geq 1 - L(2e^{-u^2/2} + e^{-u'})$. If $\sigma_k s_{v_k}$ is non-decreasing, $a_1 > \dots > a_K$, $v_k \geq 3$ ($k \geq 2$), then $t_1^{\star} < \dots < t_K^{\star}$.

Proof. Factorisation (Prop. 4.1) reduces each block to one simplex with $n = v_k$; Theorem 4.2 applies; the union bound runs over L positions. For the order: $t_k^{\star}$ increases in $\sigma_k s_{v_k}$ and decreases in a_k .

Corollary 4.6

Linear schedules with $\sigma_k s_{v_k}$ increasing commit at $t_k^{\star} = \frac{\sigma_k s_k}{1 + \sigma_k s_k}$, in order.

Theorems 4.7 and 4.9: noise memory and order



Theorem 4.7 (noise memory)

In flow time $\varsigma = \int_0^S M(\mathbf{X}_r) dr$, for every block,

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

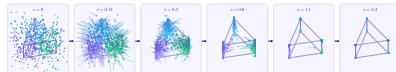
$\lambda_k = e^{-a_k \varsigma}$, $\mathbf{y}^{(\ell,k)} \in \Delta^{v_k-1}$: exactly the prior noise $\alpha_k(t) \mathbf{x}_0$ of the interpolant at $t = 1 - e^{-\varsigma}$.

Proof. Lemma 3.6 gives $\frac{d\mathbf{Z}}{d\varsigma} = a_k(\bar{\mathbf{p}} - \mathbf{Z})$; variation of constants writes $\mathbf{Z}$ as $e^{-a_k \varsigma} \mathbf{x}_0$ plus a convex average of the $\bar{\mathbf{p}}$'s.

Theorem 4.9 (order of coordinates)

If μ_1 is invariant under swapping symbols i, i' of a block, the hyperplane $x_i = x_{i'}$ is invariant and the sign of $X_i - X_{i'}$ is constant. For product-uniform data the map is the blockwise argmax, $\mathbf{T}(\mathbf{x}_0)^{(\ell,k)} = \mathbf{e}_{\arg \max_i x_{0,i}^{(\ell,k)}}$.

(a) Particles under the autonomous flow $\dot{X}_i = X_i G_i$



Proposition 4.24: energy-matched comparison



Proposition 4.24 (energy-matched comparison)

Scale level vertices by c ; the SNR of a block becomes $c\rho_k$. With $v_k = v$, $V = v^K$, equal schedules and the energy of a flat token ($c = K^{-1/2}$), the finest level crosses at $\rho = \sqrt{2K \log(v-1)}$, the flat token at $\sqrt{2 \log(V-1)}$:

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

Proof. $(v-1)^K \leq v^K - 1$ for $v \geq 2$ (expand $(v-1+1)^K$); take logarithms. For the limit, $\log(v^K - 1) = K \log v + o(1)$ and $\log(v-1) = \log v + o(1)$.

What the code changes. It adds no signal. It gives partial information earlier (coarse levels), a smaller state and, by Theorem 6.8, a map that a narrow layer can read.



5

Training and the schedule rule

Proposition 5.1: fixed points of the DBTM loss



The loss on the stacked state: $\mathcal{L}_{\text{tr}}(\theta) = \mathbb{E} \sum_{\ell,k} \left\| \mathbf{T}_{\theta}^{(\ell,k)}(\mathbf{I}_t) - \text{sg}(\mathbf{T}_{\theta}^{(\ell,k)} + \mathbf{J}_{\mathbf{T}_{\theta}^{(\ell,k)}} \dot{\mathbf{I}}_t) \right\|^2$.

Proposition 5.1 (regression step)

For fixed $C^1 \bar{\mathbf{T}}$, the minimiser over $\mathbf{T}$ 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

$$\mathbf{T}^* = \bar{\mathbf{T}} + \mathbf{J}_{\bar{\mathbf{T}}} \mathbf{b} \quad \text{a.e.,}$$

and $\bar{\mathbf{T}}$ is a fixed point iff $\mathbf{J}_{\bar{\mathbf{T}}} \mathbf{b} = 0$ off the data: the conservation equation.

Proof. An L^2 regression is solved by the conditional expectation given $\mathbf{I}_t$; $\mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{x})$ is a function of $\mathbf{x}$, so $\mathbb{E}[\mathbf{J}_{\bar{\mathbf{T}}} \dot{\mathbf{I}}_t \mid \mathbf{I}_t = \mathbf{x}] = \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{x}) \mathbb{E}[\dot{\mathbf{I}}_t \mid \mathbf{I}_t = \mathbf{x}] = \mathbf{J}_{\bar{\mathbf{T}}}(\mathbf{x}) \mathbf{b}(\mathbf{x})$.

Theorem 5.2: the training loss controls the samples



Theorem 5.2 (training loss controls sample error)

With $\varepsilon_{\text{bnd}} = \mathbb{E}_{\mathbf{y} \sim \mu_1} \|\mathbf{T}_\theta(\mathbf{y}) - \mathbf{y}\|$ and $R = \mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\|$:

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

and $R^2 \leq \mathbb{E}_q \|\mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}\|^2 = \mathcal{E}(\mathbf{T}_\theta; \mathbf{T}_\theta) - \min_{\mathbf{T}} \mathcal{E}(\mathbf{T}; \mathbf{T}_\theta) \leq \mathcal{L}_{\text{tr}}(\theta)$; for softmax outputs $\varepsilon_{\text{bnd}} \leq 2c \mathbb{E} \sum_{\ell, k} \text{CE}$, the boundary loss.

Idea. Along an exact trajectory $\frac{d}{ds} \mathbf{T}_\theta(\mathbf{X}_s) = \mathbf{J}_{\mathbf{T}_\theta} \mathbf{b}(\mathbf{X}_s)$; integrate to τ and add the boundary error at the end point. The occupation measure of the flow is at most q (Thm. 3.8(iv)).

Proposition 5.5 (anchoring)

Cross-entropy against $\mathbf{x}_1^{(\ell, k)}$ is Bayes-consistent only after the level's threshold: its target differs from the posterior by $2 \mathbb{E}[1 - \pi_\star^{(\ell, k)}]$, small for $\rho_k \geq s_{v_k} + 2u$, large below.

Training time inside the commitment window



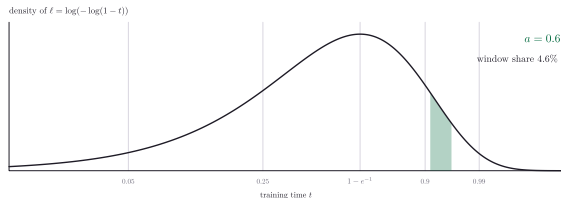
A level learns its decision only on training states inside its **commitment window**

$$|\rho(t) - s_n| \leq z.$$

In $\ell = \log(-\log(1 - t))$, uniform t is **Gumbel** distributed, and a power schedule (a, σ) :

the exponent *moves* the window,

the scale *sets its length*.



Proposition 6.2 and Theorem 6.3: the window share



Proposition 6.2 (Gumbel coordinates)

$\ell(t) = \log(-\log(1-t))$ has density $\varphi_G(\ell) = e^{\ell-e^\ell}$ for $t \sim \mathcal{U}[0, 1]$. The window $\{|\rho(t)-s| \leq z\}$ of $(1-t)^a$ with scale σ is $[c, c+\lambda]$, $c = \log \frac{\log(1+\sigma(s-z))}{a}$, $\lambda = \log \frac{\log(1+\sigma(s+z))}{\log(1+\sigma(s-z))}$, and $\sigma \mapsto \lambda$ is a decreasing bijection onto $(0, L_s)$, $L_s = \log \frac{s+z}{s-z}$: the exponent moves the window, the scale sets its length.

Theorem 6.3 (the largest window share)

For every power schedule $W(z) < M(L_s) = \exp\left(-\frac{L_s}{e^{L_s}-1}\right) - \exp\left(-\frac{L_s e^{L_s}}{e^{L_s}-1}\right)$, approached as $\sigma \rightarrow 0$; the optimal window contains $t = 1 - 1/e$. $M(L) \leq L/e$, so a level spends at most $\approx \frac{2z}{es_n}$ of training in its window; the linear schedule reaches at most $\frac{2z}{(\sqrt{s+z}+\sqrt{s-z})^2} \approx \frac{e}{4} M(L_s)$.

Corollary 6.4: the schedule rule



Corollary 6.4 (a schedule rule for DBTM)

For fixed σ , the window share is maximised exactly at

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

and then the commitment time lies in a window that contains $t = 1 - 1/e$.

DBTM's linear schedule ($a = 1, \sigma = 1$) commits later as the vocabulary grows: $t^* = 0.82$ at $V = 50,257$, with 15% less training time on the decision than $a^* = 1.70$.

Idea. In ℓ the share is the Gumbel mass of an interval of fixed length; it is largest where the density takes the same value at both ends.

Full proof of Proposition 6.2: Gumbel coordinates



1 **The law of ℓ .** $\mathbb{P}(\ell \leq l) = \mathbb{P}(-\log(1-t) \leq e^l) = \mathbb{P}(t \leq 1 - e^{-e^l}) = 1 - e^{-e^l} =: G(l)$, density $\varphi_G(l) = e^l - e^{2l}$.

2 **Schedules translate.** $\rho(t) = \frac{(1-t)^{-a}-1}{\sigma}$ increases; for $\rho_0 > 0$:

$$\rho(t) \leq \rho_0 \iff -a \log(1-t) \leq \log(1 + \sigma \rho_0)$$

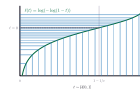
$$\iff \ell(t) \leq \log \frac{\log(1 + \sigma \rho_0)}{a}.$$

3 **The window.** $\{s-z \leq \rho \leq s+z\} = \{\ell \in [c, c+\lambda]\}$ with $c = \log \frac{\log(1 + \sigma(s-z))}{a}$, $\lambda = \log \frac{\log(1 + \sigma(s+z))}{\log(1 + \sigma(s-z))}$: a enters only c .

4 **The length.** With $q = \frac{s+z}{s-z}$, $y = \sigma(s-z)$: $\lambda = \log h(y)$, $h(y) = \frac{\log(1+qy)}{\log(1+y)} \rightarrow q$ ($y \rightarrow 0$), $\rightarrow 1$ ($y \rightarrow \infty$).

5 **h decreases.** $\psi(u) = \frac{u \log u}{u-1}$ has $\psi' = \frac{u-1-\log u}{(u-1)^2} > 0$. With $A = 1+y < B = 1+qy$, $B-1 = q(A-1)$: $qA \log A = q\psi(A)(A-1) < \psi(B)(B-1) = B \log B$, and $h'(y) = \frac{qA \log A - B \log B}{AB(\log A)^2} < 0$.

6 **Conclude.** $\sigma \mapsto \lambda$ is a decreasing bijection onto $(0, L_s)$, $L_s = \log q$: **the exponent moves the window, the prior scale sets its length.** ■



uniform t , seen in
 $\ell = \log(-\log(1-t))$

Full proof of Corollary 6.4



1 **Share.** For fixed σ the window is $[c, c + \lambda]$ with λ fixed, so the share is $f(c) = G(c + \lambda) - G(c)$.

2 **Derivative.** $f'(c) = \varphi_G(c + \lambda) - \varphi_G(c)$ has the sign of $\log \varphi_G(c + \lambda) - \log \varphi_G(c) = \lambda - e^c(e^\lambda - 1)$.

3 **Unique maximum.** This is strictly decreasing in c and vanishes at $c^* = \log \frac{\lambda}{e^\lambda - 1}$: f increases, then decreases.

4 **The value.**

$$f(c^*) = e^{-e^{c^*}} - e^{-e^{c^*} + \lambda}$$

$$= e^{-\frac{\lambda}{e^\lambda - 1}} - e^{-\frac{\lambda e^\lambda}{e^\lambda - 1}} = M(\lambda) \leq \frac{\lambda}{e},$$

, since $\varphi_G \leq \varphi_G(0) = e^{-1}$.

5 **The exponent.** Solve $c^* = \log \frac{\log(1 + \sigma(s_n - z))}{a}$:

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

6 **Location.** $e^\lambda > 1 + \lambda$ gives $e^{c^*} < 1 < e^{c^* + \lambda}$: the optimal window contains $\ell = 0$, the mode of φ_G , i.e. $t = 1 - 1/e$, for every vocabulary. ■



the window share =
Gumbel mass of an
interval

Proposition 6.5: interval packing



Proposition 6.5 (interval packing)

Fix $\sigma_k, \lambda_k = \lambda_{s_{v_k}}(\sigma_k)$. The largest $\min_k W_k(z)$ over disjoint, coarse-first windows is the largest m for which the greedy chain $c_k = \max\{e_{k-1}, \inf C_k(m)\} \leq \sup C_k(m)$, $e_k = c_k + \lambda_k$ exists, $C_k(m) = \{c : G(c + \lambda_k) - G(c) \geq m\}$; exponents $a_k = \log(1 + \sigma_k(s_{v_k} - z))e^{-c_k}$, and $m^\star \leq \min\{\min_k M(\lambda_k), 1/K\}$.

V	flat	K = 2	K = 3
256	0.188	0.286	0.222
1,024	0.161	0.247	0.229
4,096	0.142	0.219	0.221
16,384	0.128	0.198	0.211
50,257	0.119	0.184	0.202

Window share of every level under the best ordered exponents ($\sigma = 0.55, z = 1.28$).

Theorem 6.7: the embedding bottleneck



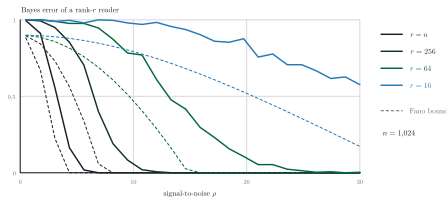
A network reads each token's state through a linear layer of rank $r \ll V$.

Theorem 6.7 (the bottleneck delays commitment)

Every estimator of the symbol from $W\mathbf{I}_t$ errs with probability

$$\geq 1 - \frac{\frac{r}{2} \log(1 + \rho^2/n) + \log 2}{\log n}.$$

So it commits only at $\rho \approx s_n \sqrt{n/r}$, polynomial in n . A coarse-to-fine state with $\sum_k v_k \leq r$ is read without loss.



Proof: Fano's inequality and the capacity of an r -dimensional Gaussian channel.

Full proof of Theorem 6.7



1 **Three tools.** Fano: $H(V | \hat{V}) \leq \log 2 + P_e \log(n-1)$.
Data processing: $I(V; Y') \leq I(V; Y)$ along $V \rightarrow Y \rightarrow Y'$. Gaussian max. entropy: $h(Z) \leq \frac{1}{2} \log((2\pi e)^r \det K)$.

2 **r coordinates.** Let $U \in \mathbb{R}^{r \times n}$ have orthonormal rows spanning the rows of W ; then $W = WU^T U$ and $Ux = L_W x$ for a left inverse L . So $\hat{v}$ is a function of $Z = \frac{U I_t}{\alpha \sigma} = \rho u_{i^*} + \zeta$, $\zeta = U g \sim \mathcal{N}(0, I_r)$.

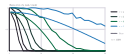
3 **Fano.** $H(i^*) = \log n$, so
$$P_e \geq 1 - \frac{I(i^*; \hat{i}) + \log 2}{\log n}.$$

4 **Data processing.** $i^* \rightarrow X = \rho u_{i^*} \rightarrow Z \rightarrow \hat{i}$ is Markov, so $I(i^*; \hat{i}) \leq I(X; Z) = h(Z) - h(\zeta) \leq \frac{1}{2} \log \det(I + \text{Cov} X)$.

5 **AM-GM.** $\det(I + \text{Cov} X) \leq (1 + \frac{\text{tr Cov} X}{r})^r$ and $\text{tr Cov} X \leq \mathbb{E} \|X\|^2 = \frac{\rho^2}{n} \sum_i \|U e_i\|^2 = \frac{\rho^2 r}{n}$.

6 **Conclude.** $I \leq \frac{r}{2} \log(1 + \rho^2/n)$, hence
$$P_e \geq 1 - \frac{\frac{r}{2} \log(1 + \rho^2/n) + \log 2}{\log n};$$

error $\leq \varepsilon$ needs $\rho^2 \geq \frac{2n}{r} ((1 - \varepsilon) \log n - \log 2)$. ■



error of any reader of a

rank- r layer

Theorem 6.8: the bottleneck hides the transport map



Theorem 6.8 (the bottleneck hides the transport map)

Let $K = 1$, μ_1 uniform on n vertices, so $T(x_0) = e_{\arg \max_i x_{0,i}}$ (Thm. 4.9). For $\hat{i}$ the vertex nearest to $f(Wx_0)$, W of rank r :

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

$\mathbb{E}\|f(Wx_0) - T(x_0)\| \geq (1 - p_n(r))/\sqrt{2}$, and $p_n(r) \leq 2r \log n/n + o(1)$.

Idea. The projection onto the row space of W has trace r : at most r/θ coordinates keep weight $\geq \theta$, while the argmax of the prior draw is uniform over all n .

Corollary 6.9 (a floor for the flat objective)

Every flat $T_\theta = f \circ W$ has $\varepsilon_{\text{bnd}} + R \geq (1 - p_n(r))/\sqrt{2}$; a coarse-to-fine state read injectively has no such floor.



6 *Experiments*

Experiment: the schedule rule in DBTM



Synthetic language with $V = 1,024$ tokens (exact likelihood), the same network for both schedules.

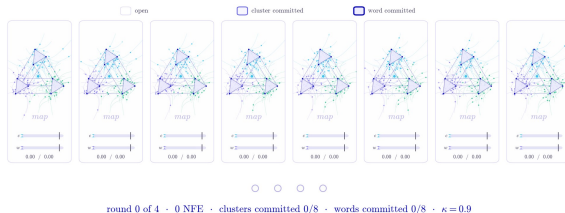
schedule	standard training (3 seeds)			refinement in the loop		
	trans. $k=1$	trans. $k=4$	NLL $k=4$	trans. $k=4$	trans. $k=8$	seq. $k=8$
linear, $a = 1$	16.5	31.2	9.47	56.5	88.2	51.1
$a^\star = 1.50$	20.6	39.1	8.58	59.5	94.3	72.2

One scalar changes; the network, losses and sampler stay the same. With refinement-in-loop training the rule removes 43% of the invalid sequences.

Experiment: coarse-to-fine and flat DBTM



- One-step samples are equally valid (15.8% vs 16.0%).
- The coarse level is much better: 71% of cluster transitions valid, against 29%.
- Given true context, it fills gaps better: 21.5% vs 15.1% exact tokens.
- It loses under refinement; committing whole tokens and training in the loop close the gap at four evaluations.



refinement with exact per-level posteriors: clusters commit first, then words

Sudoku: refinement rounds



confident
 unsure
 wrong

1	8	7	3	9	2	4	6	6
5	9	4	5	6	1	8	7	3
2	6	3	8	4	7	5	9	1
8	2	1	6	3	4	7	5	9
7	4	9	3	1	5	6	3	8
6	3	5	7	9	8	1	4	2
4	1	6	2	8	9	3	5	7
9	5	8	4	7	3	2	1	6
3	7	2	1	5	6	9	8	4

round 1

1	9	7	8	5	2	4	2	6
5	5	4	3	6	9	8	7	3
1	6	3	8	4	7	5	9	6
8	2	1	6	3	4	7	5	9
7	4	9	1	2	5	6	3	8
6	3	5	7	9	8	1	4	2
4	1	2	9	8	6	3	6	7
9	5	8	4	7	3	2	1	5
3	7	8	5	5	6	9	8	4

round 2

6	9	7	8	5	1	4	2	6
5	5	4	9	6	2	8	7	1
1	6	3	2	4	7	5	9	1
8	2	1	6	3	4	7	5	9
7	4	9	1	2	5	6	3	8
6	3	6	7	9	8	1	4	2
4	1	5	9	8	2	3	6	7
9	6	8	4	7	3	2	1	5
3	7	2	1	1	6	9	8	4

round 3

6	9	7	8	5	1	4	2	3
2	5	4	3	6	9	8	7	1
1	8	3	2	4	7	5	9	6
8	2	1	6	3	4	7	5	9
7	4	9	1	2	5	6	3	8
5	3	6	7	9	8	1	4	2
4	1	5	9	8	2	3	6	7
9	6	8	4	7	3	2	1	5
3	7	2	5	1	6	9	8	4

round 4

TinyGSM: refinement rounds



■ committed ■ refining

Charmaine will be 16 years old in 12 years. How old will she be 4 years from now?

```
def simple_math_problem() -> int:
    """
    Charmaine will be 16 years old in 12 years.
    How old will she be 4 years from now?
    """
```

```
age_age = 16
_ = _____
```

round 1

```
age = 16161 1 1 == =
result = result result
result result
```

round 10

```
x = 16 - =24
y = x + 4

result = y

return result
<|endoftext|>
```

round 20

```
x = 16 - 12
y = x + 4
result = y
return result
<|endoftext|>
```

round 31



Theory

On products of simplices the field is explicit, the flow converges for every schedule, and commitment has a sharp threshold with a universal Gaussian profile; the flow keeps the order of the levels.

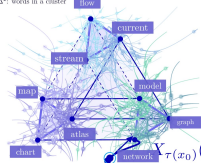
Practice

Gumbel coordinates of training time give a closed-form schedule that improves DBTM. A coarse-to-fine state wins on the coarse decision and on context; refinement is where it still loses.

state space $\Delta^2 \times \Delta^2$ (9 words)

Δ : cluster

Δ^2 : words in a cluster

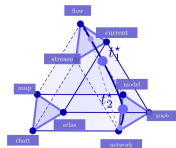


Coarse-to-Fine Map $T : \prod_{k=1}^K (\Delta^{v_k-1})^L \rightarrow M_1$

$$t_k^* = 1 - (1 + \sigma \sqrt{2 \log v_k})^{-1/a_k}$$

$$X_0(x_0) = x_0 \quad X_{\tau(x_0)}(x_0) = x_1$$

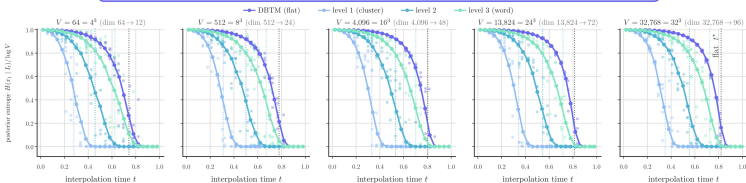
$$X_0(x_0) = x_0 \quad b(x) \cdot \nabla X_t(x) = \partial_t X_t(x)$$



$$I_t^{(k)} = \alpha_t^{(k)} x_0^{(k)} + \beta_t^{(k)} x_1^{(k)}$$

$$T(I_t) = x_1$$

$$T = \arg \min \mathbb{E}_{t, x_0, x_1} \sum_{k=1}^K \left[\|T_\theta^{(k)}(I_t) - \text{sg}(T_\theta^{(k)}(I_t) + \dot{I}_t \cdot \nabla T_\theta^{(k)}(I_t))\|^2 + \mathbf{1}[t > t_k^*] \text{CE}(T_\theta^{(k)}(I_t), x_1^{(k)}) \right]$$



Thank you.

Johannes Herter