Advances in probability and statistics using AI methods and agents

Aleksandr Zimin

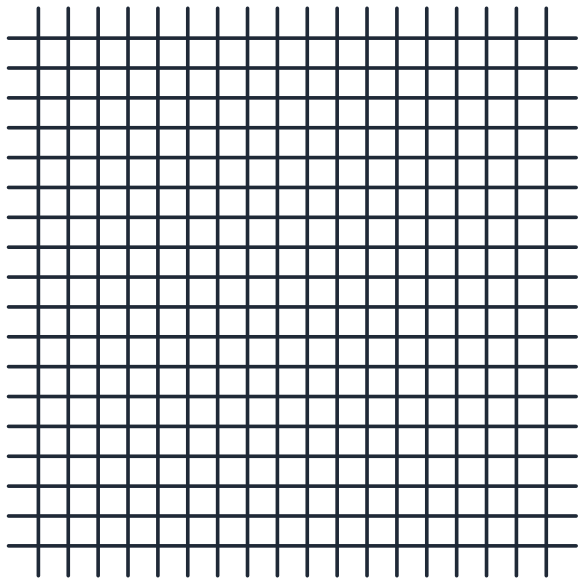
MIT Mathematics · PhD Defense

PART I

Percolation

A random maze

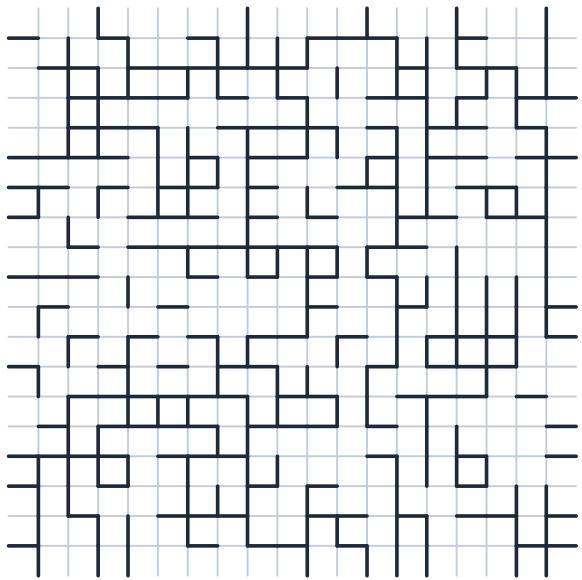
Is there a path from top to bottom?



Each edge: open with probability p, closed with $1-p$.

A random maze

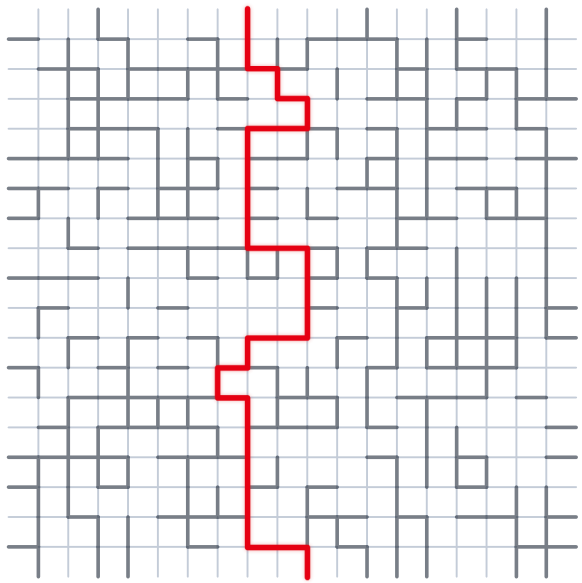
Is there a path from top to bottom?



$p = 0.5$

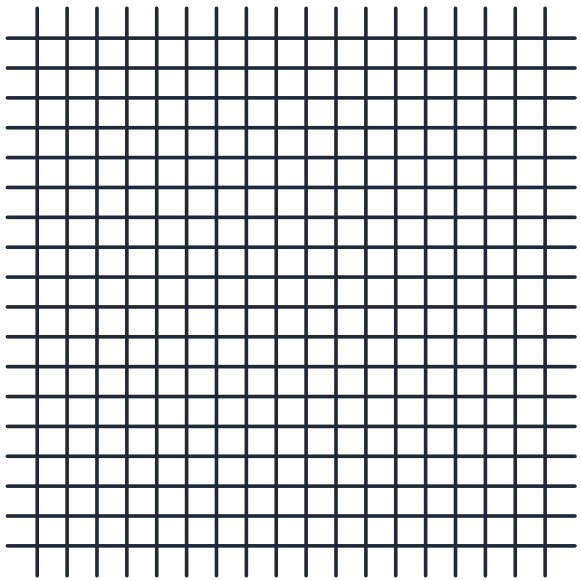
A random maze

Is there a path from top to bottom?

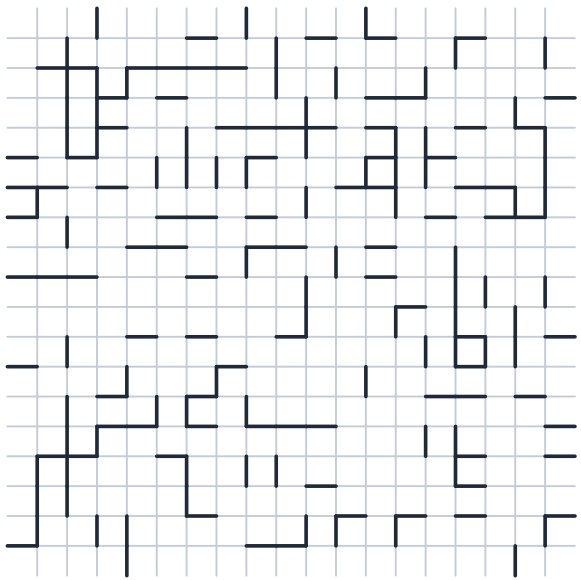


There is a path.

Critical probability

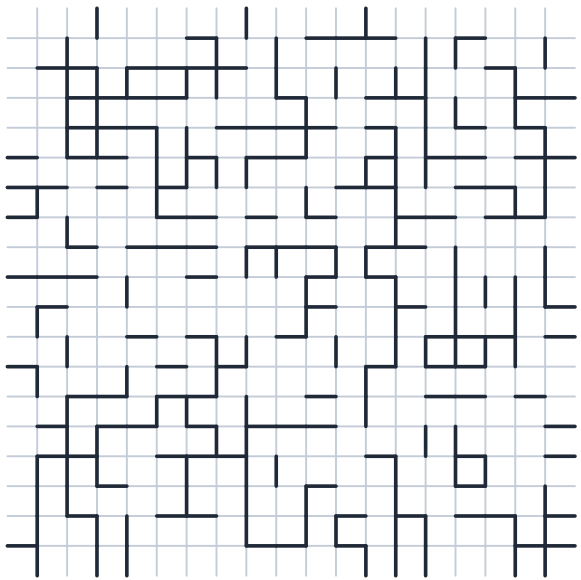


Critical probability



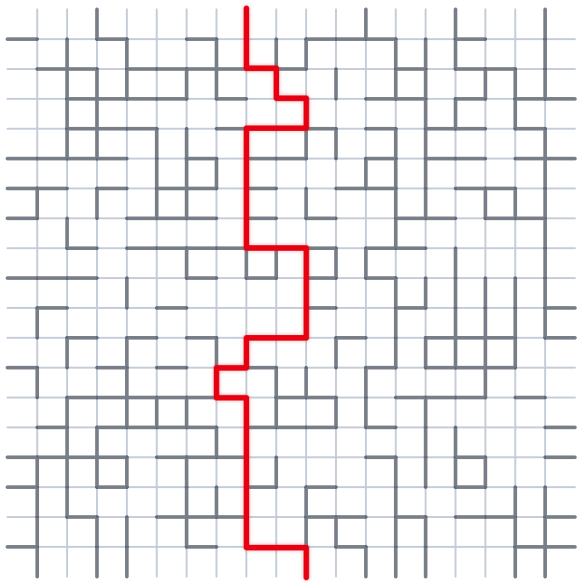
$p = 0.25$

Critical probability



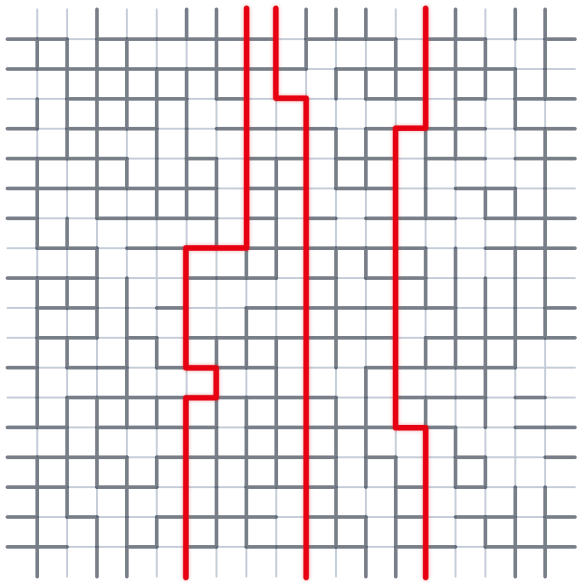
$p = 0.4$

Critical probability



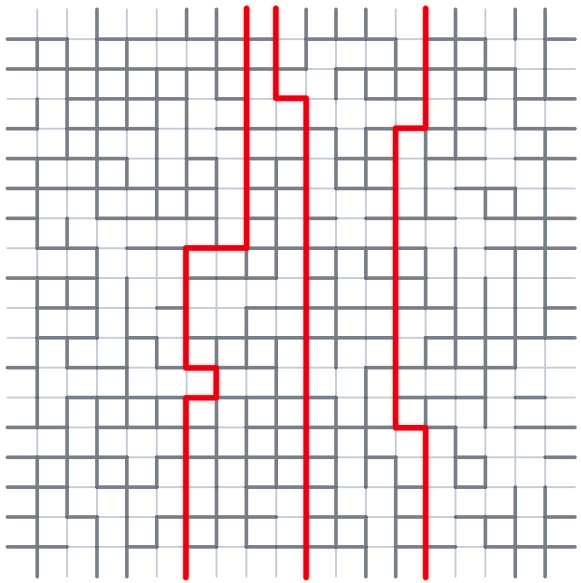
critical probability

Critical probability



$p = 0.7$ — three of many.

Critical probability

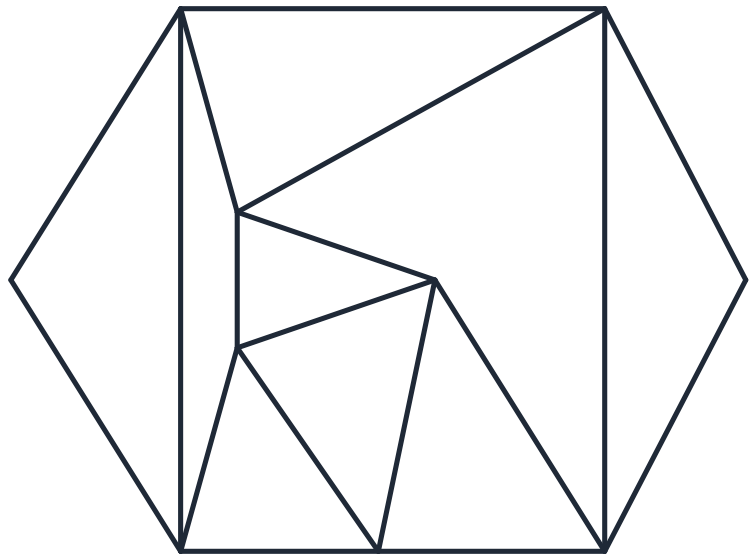


$p = 0.7$ — three of many.

Theorem (Harris 1960, Kesten 1980)

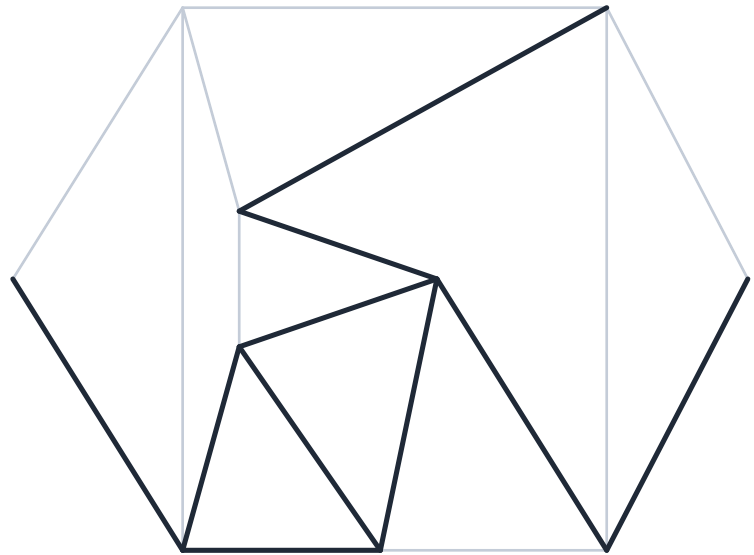
For $p \le \frac{1}{2}$, no infinite cluster on $\mathbb{Z}^2$ a.s. For $p > \frac{1}{2}$, an infinite cluster a.s.

Percolation on any graph



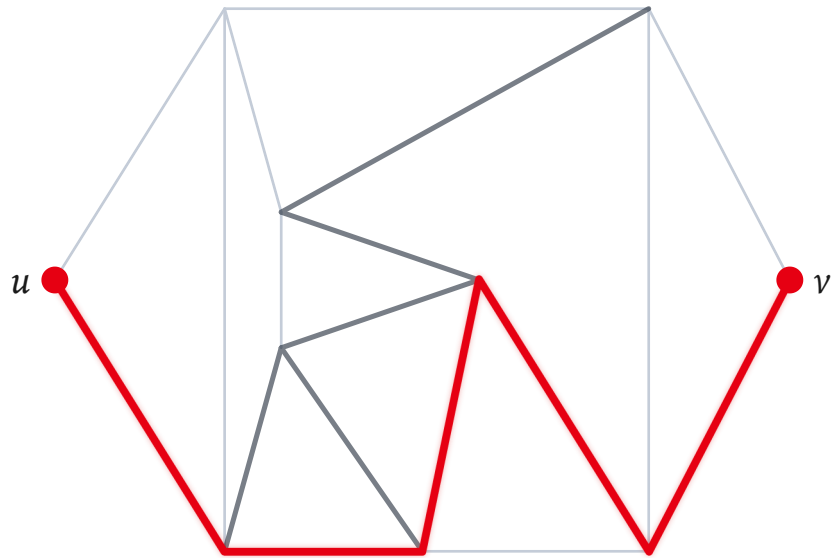
An example graph — vertices, edges.

Percolation on any graph



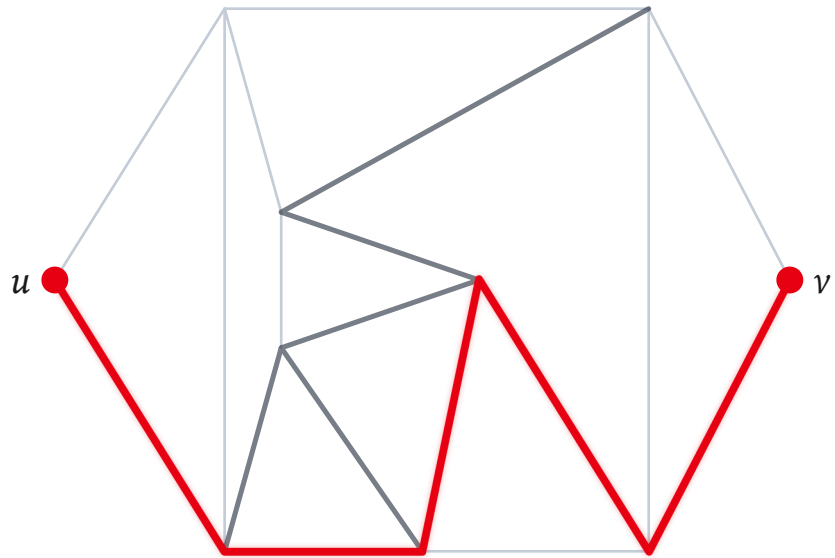
Keep each edge with probability p.

Percolation on any graph



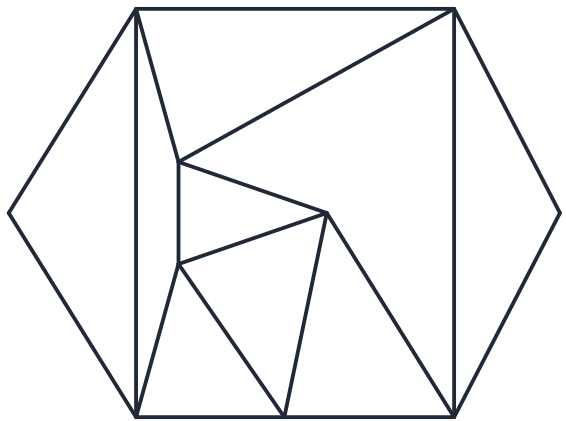
Pick two vertices — are they connected?

Percolation on any graph



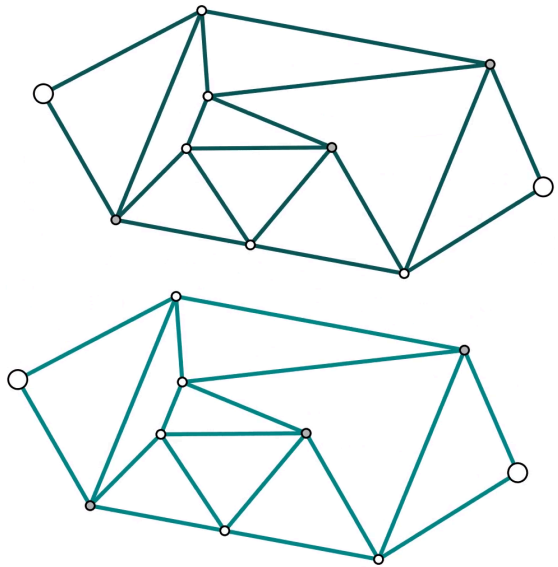
$\mathbb{P}_p(u \leftrightarrow v)$ — the connectivity probability.

Bunkbeds



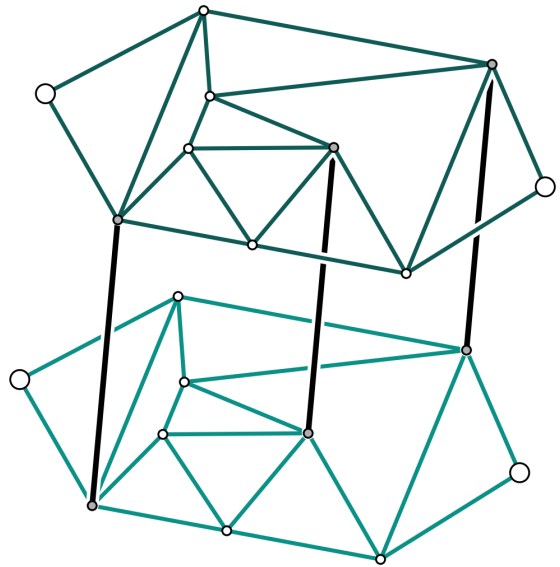
Start with a graph.

Bunkbeds



Duplicate it into two layers.

Bunkbeds

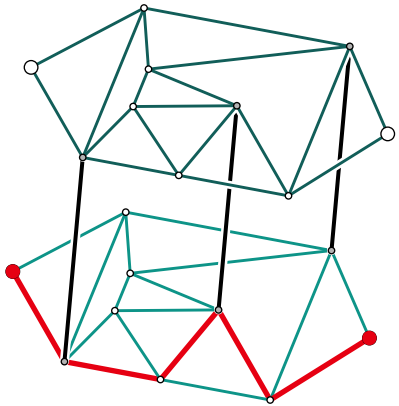


Add posts.

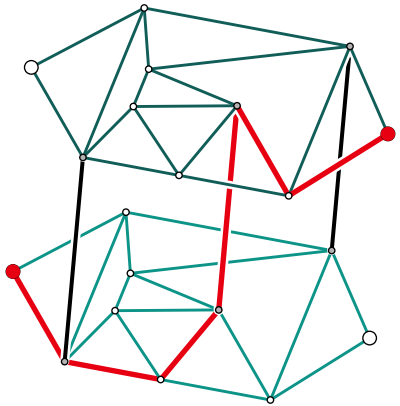
Bunkbeds

Yes — like the dorm bunkbed.

Same floor vs across floors



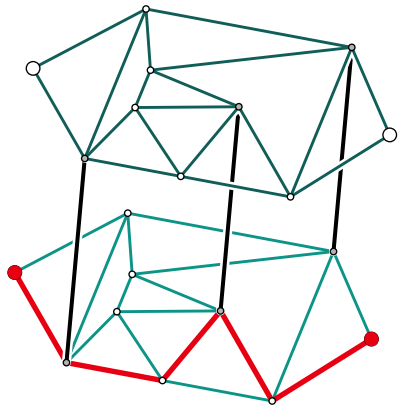
$u \leftrightarrow v$ (same floor)



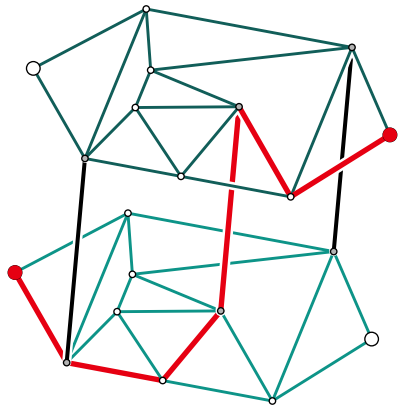
$u \leftrightarrow v'$ (across floors)

Which connection is easier?

Same floor vs across floors



$u \leftrightarrow v$ (same floor)



$u \leftrightarrow v'$ (across floors)

Conjecture (Kasteleyn, 1985)

For any connected G, any transversal T, and any $p \in (0,1)$:

$$\mathbb{P}_p^{\mathrm{bb}}(u \leftrightarrow v) \;\geq\; \mathbb{P}_p^{\mathrm{bb}}(u \leftrightarrow v')$$

Why it seemed true

Proved cases

Wheels · complete graphs · complete bipartite graphs · symmetry cases.

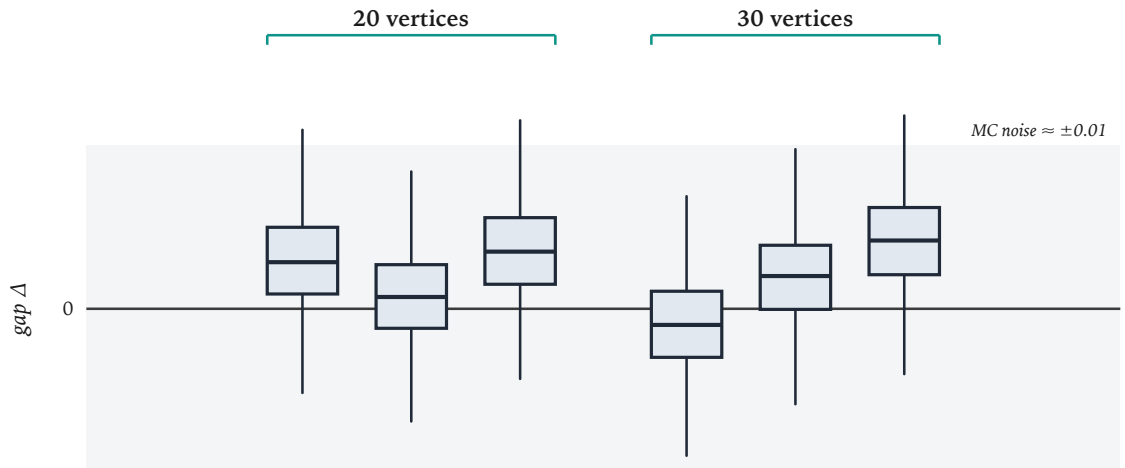
Small transversal sets: $|T| = 1$ and $|T| = 2$.

Limit regime: $p \uparrow 1$ (any graph).

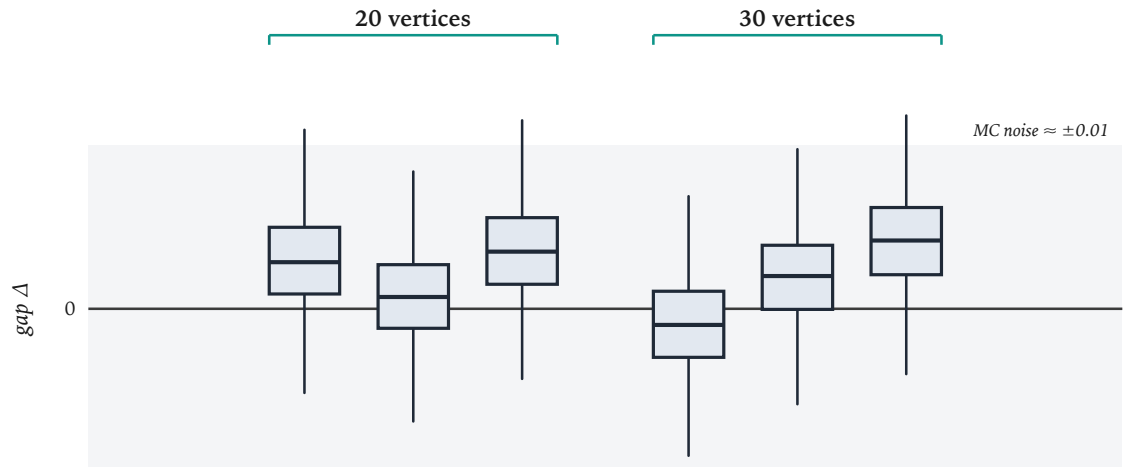
Sibling models

Simple random walks and the Ising model (Häggström, 1998 · 2003).

Monte Carlo: searching for a counterexample

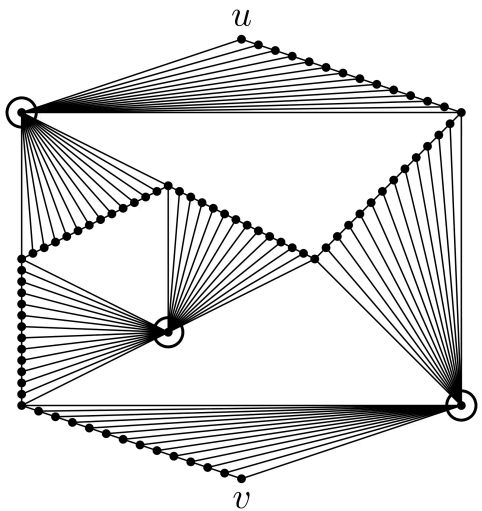


Monte Carlo: searching for a counterexample



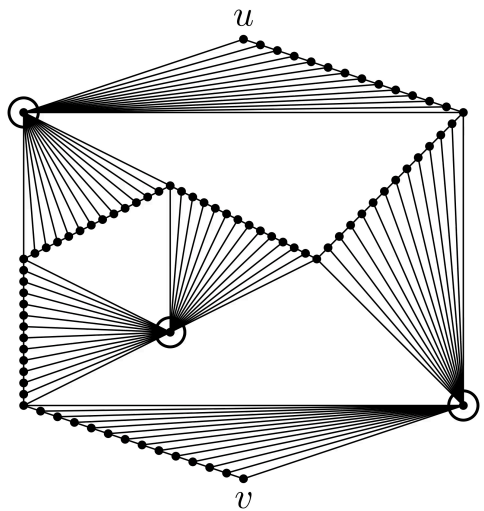
≈12,000 generated graphs · no counterexample found.

But it's false.



82-VERTEX COUNTEREXAMPLE · COMPUTER-ASSISTED

But it's false.



82-VERTEX COUNTEREXAMPLE · COMPUTER-ASSISTED

Theorem (Gladkov–Pak–Zimin, PNAS 2025)

There is a connected planar graph $G = (V, E)$ with $|V| = 7222$ vertices and $|E| = 14442$ edges, a subset $T \subset V$ with three transversal vertices, and $u, v \in V$, such that

$$\mathbb{P}^{\mathrm{bb}}_{1/2}(u \leftrightarrow v) \;<\; \mathbb{P}^{\mathrm{bb}}_{1/2}(u \leftrightarrow v').$$

The gap is far below what sampling can detect.

Too small to see.

$$10^{-47}$$

82-VERTEX COUNTEREXAMPLE · COMPUTER-ASSISTED

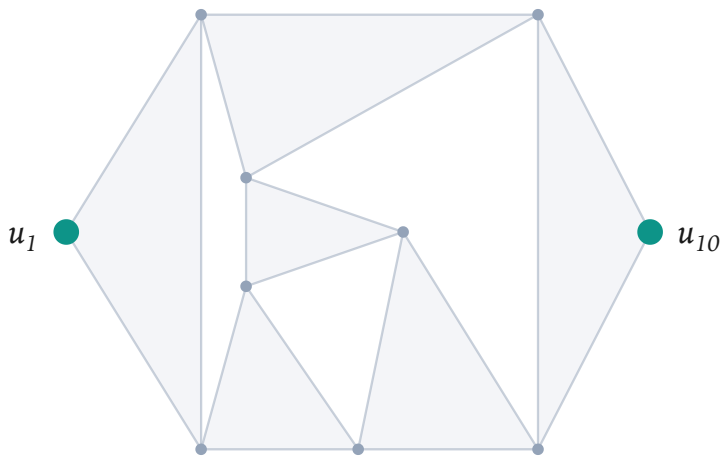
Too small to see.

$$10^{-47}$$

82-VERTEX COUNTEREXAMPLE · COMPUTER-ASSISTED

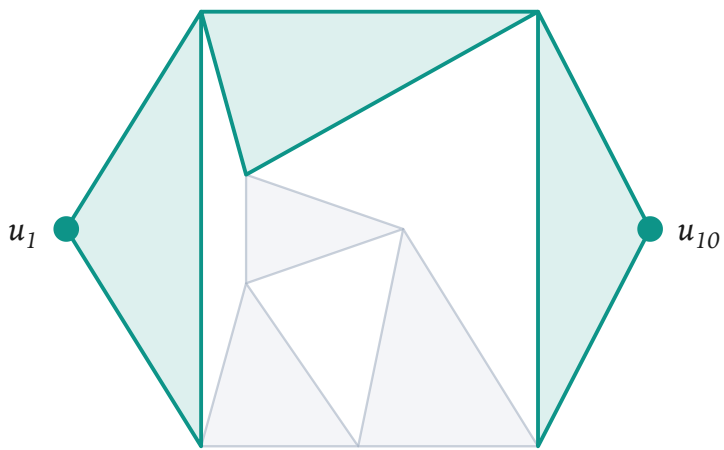
$\sim 10^{-2}$	*Monte Carlo noise floor in our runs.*
$N \sim 10^{94}$	*samples needed to resolve the gap by sampling.*
$< 10^{-4331}$	*proved 7222-vertex planar instance (PNAS 2025).*

Hollom's hypergraph counterexample



Each triangle is **one hyperedge** (not three edges).

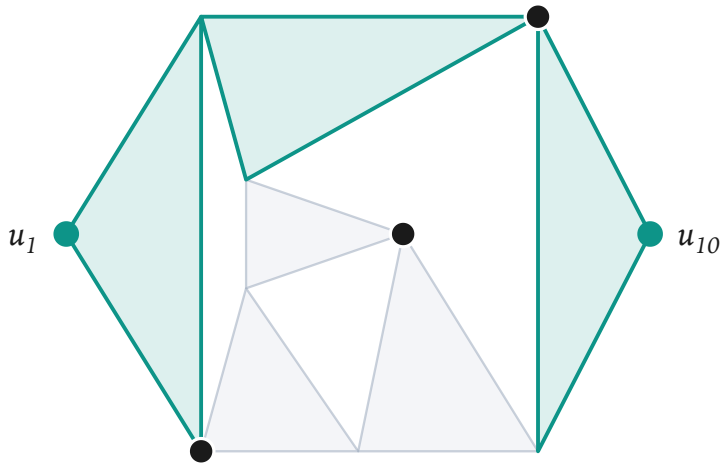
Hollom's hypergraph counterexample



Each triangle is **one hyperedge** (not three edges).

A hyperpath is a chain of hyperedges sharing a vertex.

Hollom's hypergraph counterexample



Each triangle is **one hyperedge** (not three edges).

A hyperpath is a chain of hyperedges sharing a vertex.

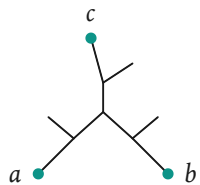
Lemma (Hollom, 2024)

In the alternative bunkbed hypergraph percolation on H with transversal set $T = \{u_2, u_7, u_9\}$:

$$\mathbb{P}(u_1 \leftrightarrow u'_{10}) = \tfrac{13}{64}$$
$$\mathbb{P}(u_1 \leftrightarrow u_{10}) = \tfrac{12}{64}$$

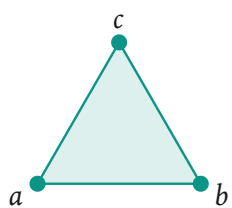
Violates the bunkbed inequality.

Can bonds simulate a 3-hyperedge?



BOND GADGET · BOND GRAPH

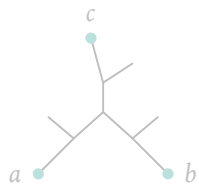
?



3-HYPEREDGE · $p = ½$

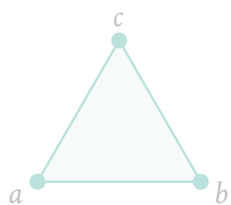
connected ½ · separate ½

Can bonds simulate a 3-hyperedge?



BOND GADGET · BOND GRAPH

×



3-HYPEREDGE · $p = ½$

connected ½ · separate ½

Theorem (Gladkov–Zimin, 2026)

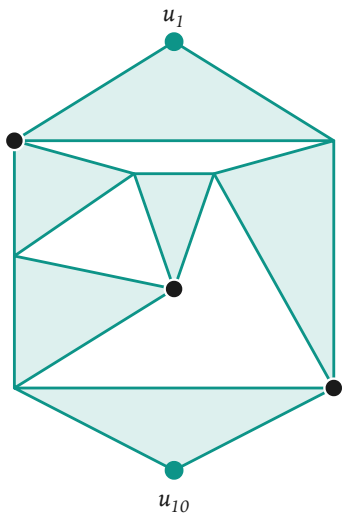
$$\mathbb{P}(abc) := \text{all three in one component},$$

$$\mathbb{P}(a \mid b \mid c) := \text{three separate components}.$$

$$\mathbb{P}(abc)\,\mathbb{P}(a \mid b \mid c) \;\le\; 1 - \mathbb{P}(abc) - \mathbb{P}(a \mid b \mid c)$$

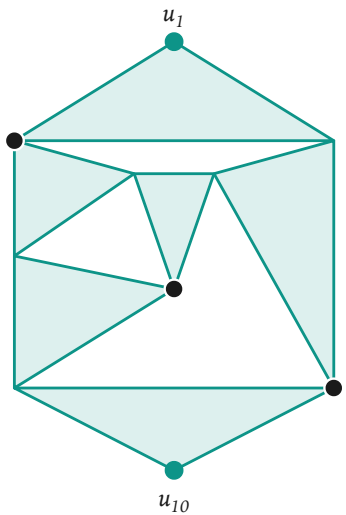
Exact simulation is out — we need a different gadget condition.

Hollom → gadget → counterexample



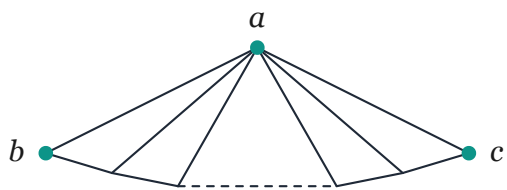
Hollom hypergraph H

Hollom → gadget → counterexample



Hollom hypergraph H

substitute each 3-hyperedge

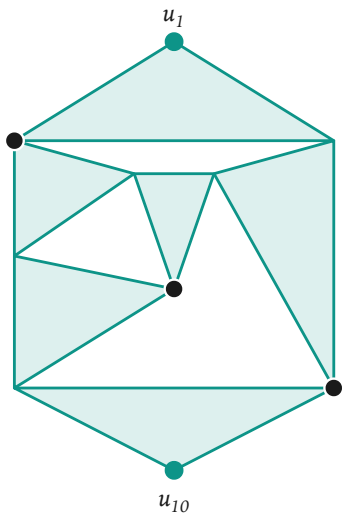


$$p = \tfrac{1}{2}, \quad n = 1204$$

Lemma (Covariance amplification gadget)

$$\mathrm{Cov}\big(a \leftrightarrow b,\ a \leftrightarrow c\big) \ > \ \Big(\tfrac{n}{3} - 1\Big)\, \mathbb{P}\big(b \leftrightarrow c \not\leftrightarrow a\big)$$

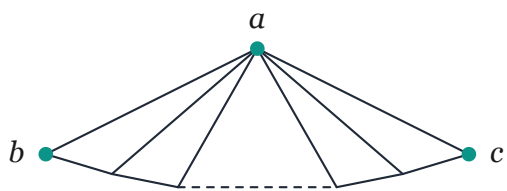
Hollom → gadget → counterexample



Hollom hypergraph H

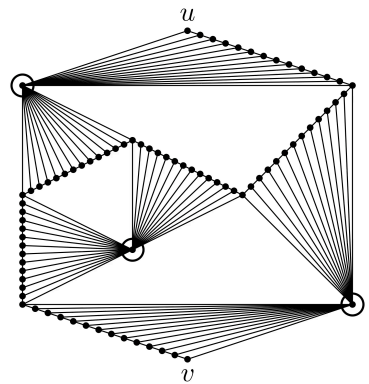
→

substitute each 3-hyperedge



$$p = \tfrac{1}{2}, \quad n = 1204$$

→



7222 vertices · 14442 edges
(illustrated here with $n = 14$)

Lemma (Covariance amplification gadget)

$$\mathrm{Cov}\big(a \leftrightarrow b,\ a \leftrightarrow c\big) \ > \ \Big(\tfrac{n}{3} - 1\Big)\, \mathbb{P}\big(b \leftrightarrow c \not\leftrightarrow a\big)$$

Open questions

Open questions

Conjecture (log-stability near the FKG equality case)

$$\mu(a \mid bc) < \varepsilon \implies \operatorname{Cov}(a \leftrightarrow b,\ a \leftrightarrow c) = O\big(\varepsilon \log(1/\varepsilon)\big).$$

Open questions

Conjecture (log-stability near the FKG equality case)

$$\mu(a \mid bc) < \varepsilon \implies \operatorname{Cov}(a \leftrightarrow b,\ a \leftrightarrow c) = O\big(\varepsilon \log(1/\varepsilon)\big).$$

Conjecture (Kozma–Nitzan, 2024)

$$\mathbb{P}(a \leftrightarrow b) \ \geq\ \mathbb{P}\big((a \leftrightarrow c_1) \cup \cdots \cup (a \leftrightarrow c_n)\big) \cdot \min_i \mathbb{P}(c_i \leftrightarrow b).$$

If true: no infinite cluster at criticality on $\mathbb{Z}^d$ (interesting: $d = 3, \ldots, 10$).

PART II

Transformers

Transformers Became Infrastructure

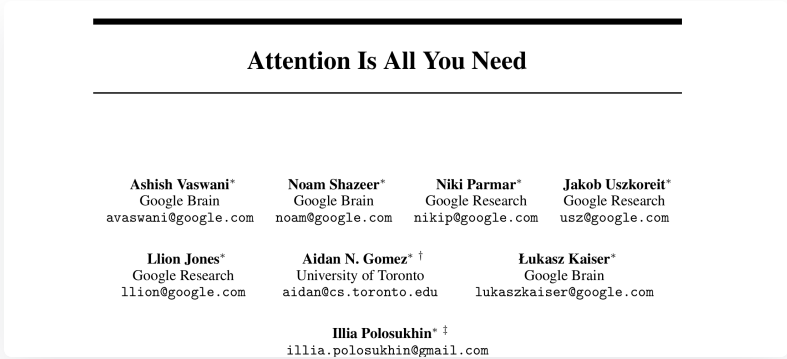
One block became a reusable backbone.

Transformers Became Infrastructure

One block became a reusable backbone.

Same block, different domains.

Dominance Showed Up in Paper Titles



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

That cultural spillover is a hint: one mechanism was quietly doing the heavy lifting.

Attention is **all you need** in speech separation
C Subakan, M Ravanelli, S Cornell… - ICASSP 2021-2021 …, 2021 - ieeexplore.ieee.org
… Transformers are emerging as a natural alternative to standard RNNs, replacing recurrent computations with a multi-head **attention** mechanism.In this paper, **we** propose the SepFormer…
☆ Save 99 Cite Cited by 1006 Related articles All 6 versions

Not **all attention** is **all you need**
H Wu, H Zhao, M Zhang - arXiv preprint arXiv:2104.04692, 2021 - arxiv.org
… is less explored, which is our **focus** in this paper. For scores dropout, **we need** to pay **attention** to a special case, when **all** attentions are shut down, that is, **all** elements in M equal to −inf …
☆ Save 99 Cite Cited by 14 Related articles All 4 versions »

Is **Attention All You Need**?
P Mineault - From Human **Attention** to Computational **Attention**: **A** …, 2025 - Springer
… **We** show how transformers and **attention have** opened up new dimensions of inquiry in … [10], who first introduced **what they** called intra-**attention**, **what we** now call self-**attention**. …
☆ Save 99 Cite Cited by 5 Related articles All 3 versions

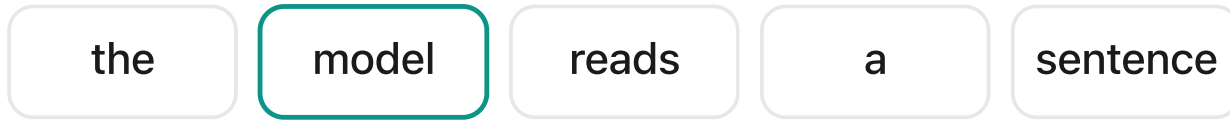
Attention is not **all you need** anymore
Z Chen - arXiv preprint arXiv:2308.07661, 2023 - arxiv.org
… and memory complexity or perform close to or better with less computational and memory complexity, confirming that the self-**attention** mechanism is indeed not **all we need** anymore. …
☆ Save 99 Cite Cited by 7 Related articles All 2 versions »

Channel **attention** is **all you need** for video frame interpolation
M Choi, H Kim, B Han, N Xu, KM Lee - … of the AAAI conference on artificial …, 2020 - aaai.org
… Since every spatial region is important for pixel-level synthesis, **we** do not explicitly consider spatial **attention**. Instead, **we** adopt the channel **attention** method proposed in (Zhang et al. …
☆ Save 99 Cite Cited by 422 Related articles All 9 versions »

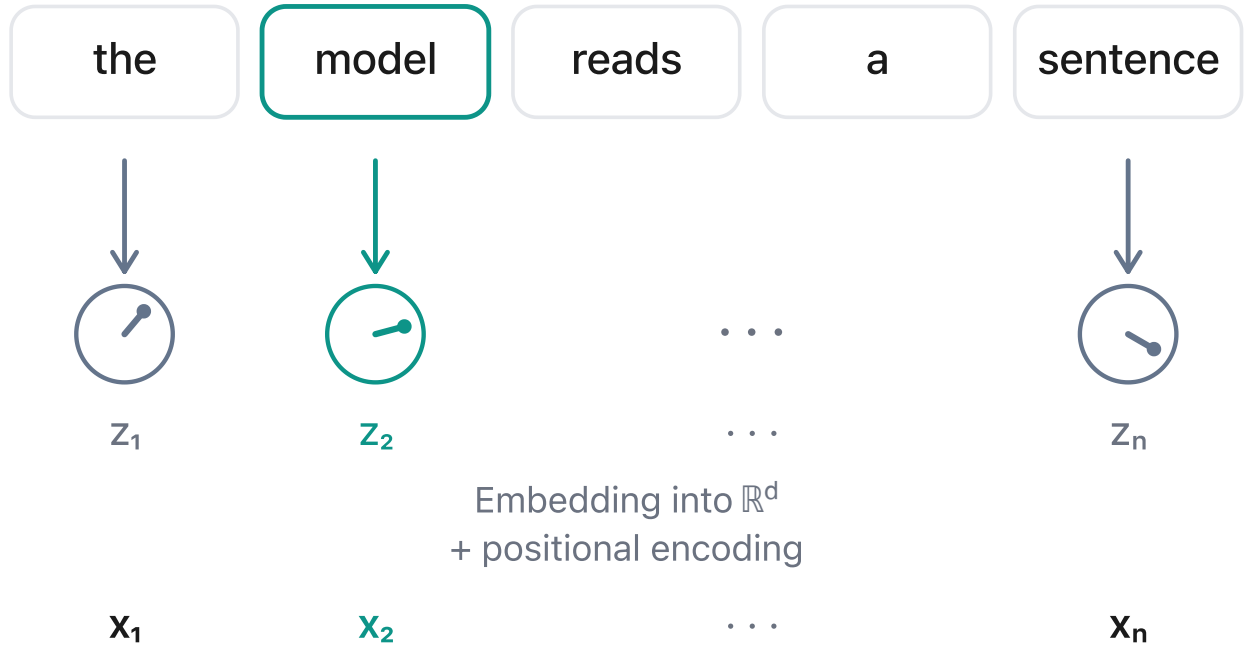
Prompt → Tokens → Vectors

" the model reads a sentence"

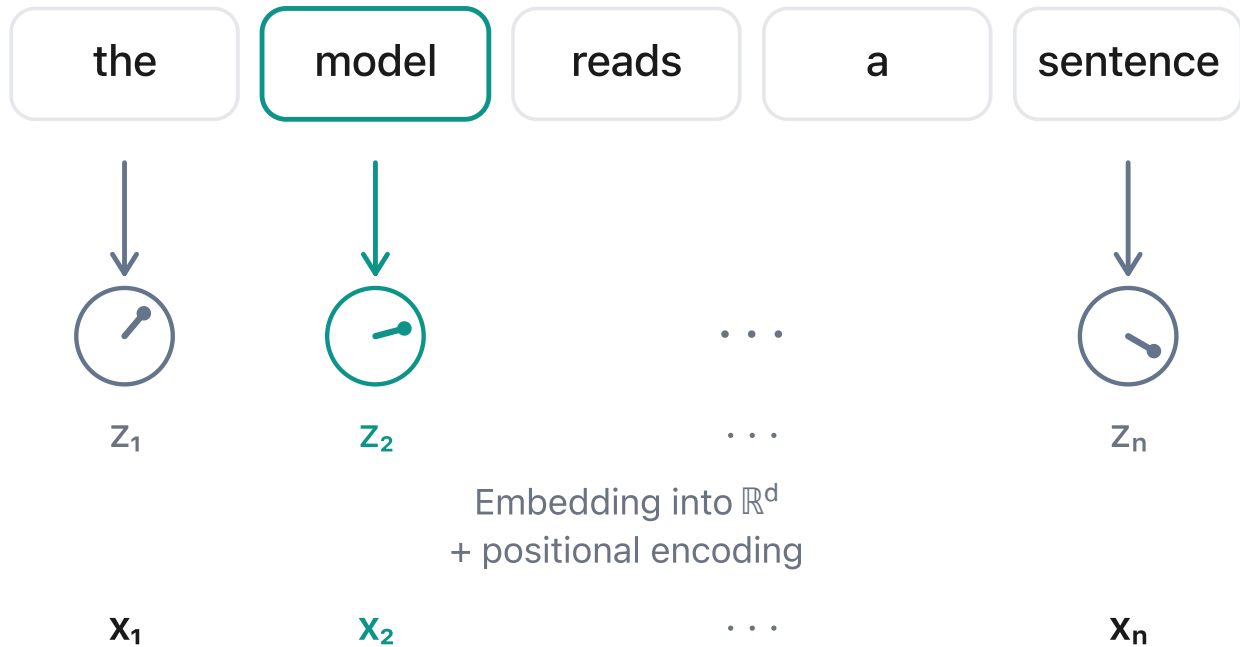
Prompt → Tokens → Vectors



Prompt → Tokens → Vectors

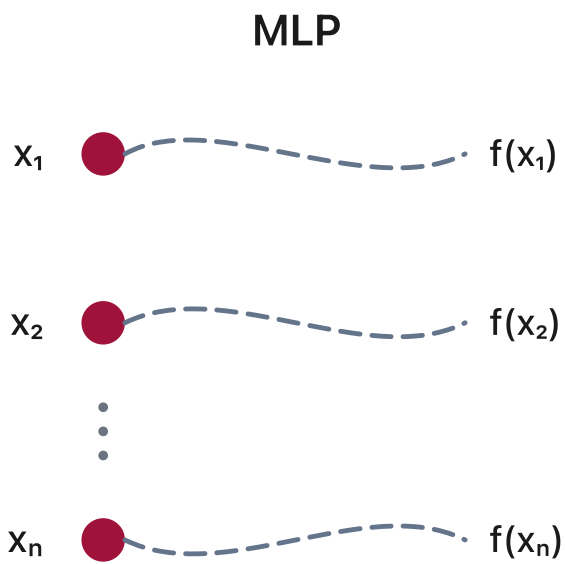


Prompt → Tokens → Vectors

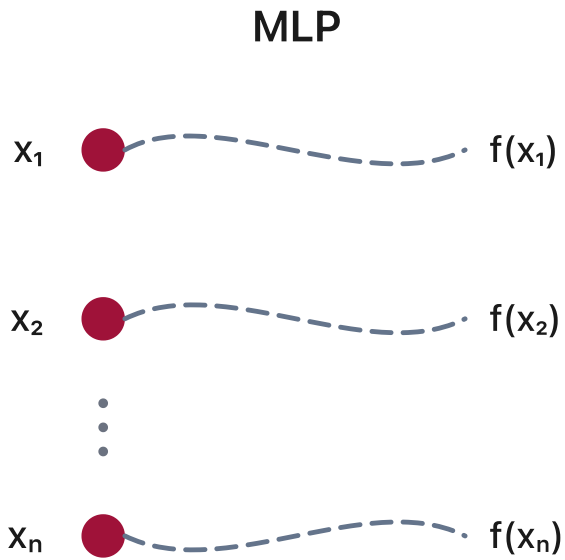


$x_1, \ldots, x_n \in \mathbb{R}^d$: tokens — the state is a cloud of vectors

Transformers Are Not Pointwise Networks

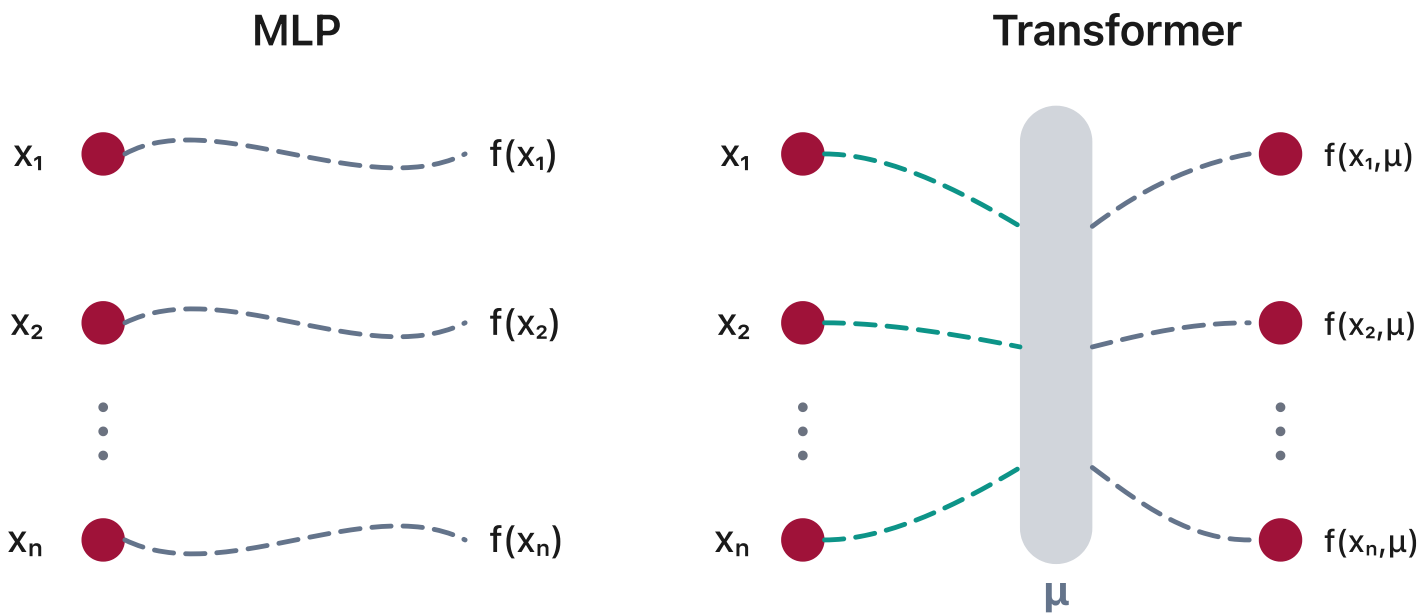


Transformers Are Not Pointwise Networks



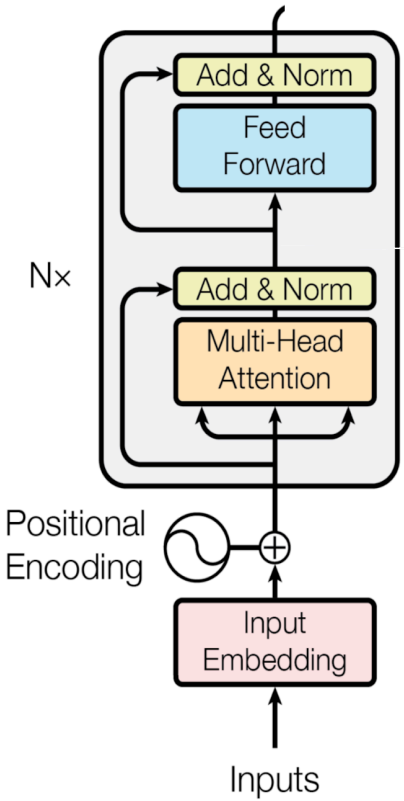
Empirical measure $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

Transformers Are Not Pointwise Networks



Empirical measure $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

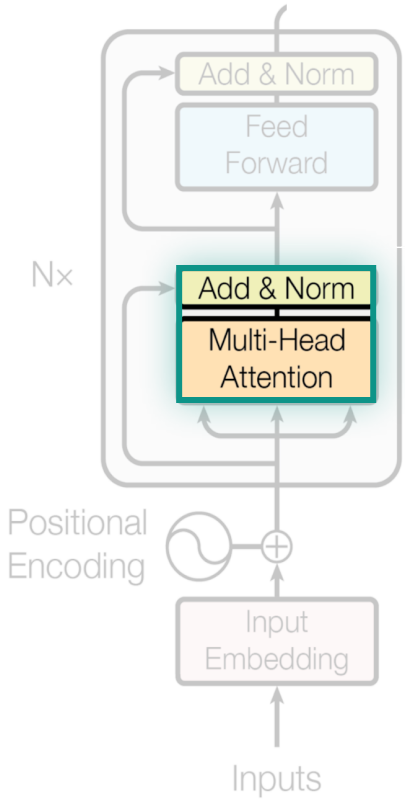
One Transformer Layer



input tokens

$x_1, x_2, \ldots, x_n \in \mathbb{R}^d$

One Transformer Layer



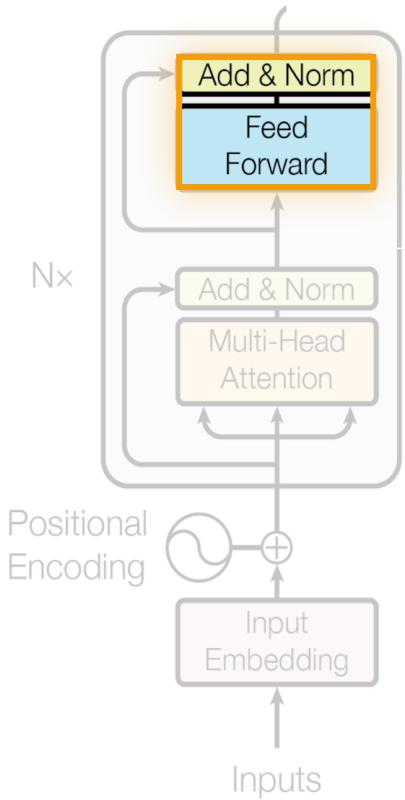
attention: mix across tokens

$$x_i \mapsto x_i + \sum_{j=1}^{n} \frac{\exp\big(\beta \langle Q x_i,\ K x_j \rangle\big)}{\sum_{k=1}^{n} \exp\big(\beta \langle Q x_i,\ K x_k \rangle\big)} V x_j$$

input tokens

$$x_1, x_2, \ldots, x_n \in \mathbb{R}^d$$

One Transformer Layer



MLP: reshape each token

$$x_i \;\mapsto\; x_i + \sigma(Ax_i + b)$$

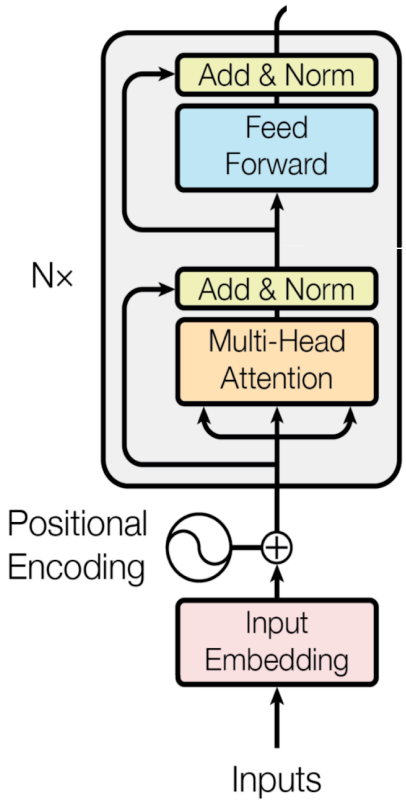
attention: mix across tokens

$$x_i \;\mapsto\; x_i \;+\; \sum_{j=1}^{n} \frac{\exp\big(\beta\,\langle Qx_i,\; Kx_j\rangle\big)}{\sum_{k=1}^{n} \exp\big(\beta\,\langle Qx_i,\; Kx_k\rangle\big)}\, Vx_j$$

input tokens

$$x_1, x_2, \ldots, x_n \in \mathbb{R}^d$$

One Transformer Layer



output tokens

$$y_1, y_2, \ldots, y_n \in \mathbb{R}^d$$

MLP: reshape each token

$$x_i \;\mapsto\; x_i + \sigma(Ax_i + b)$$

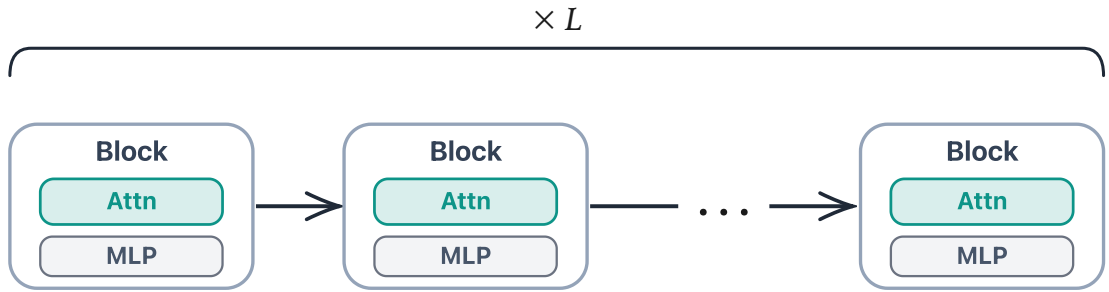
attention: mix across tokens

$$x_i \;\mapsto\; x_i \;+\; \sum_{j=1}^{n} \frac{\exp\big(\beta\,\langle Qx_i,\; Kx_j\rangle\big)}{\sum_{k=1}^{n} \exp\big(\beta\,\langle Qx_i,\; Kx_k\rangle\big)}\, Vx_j$$

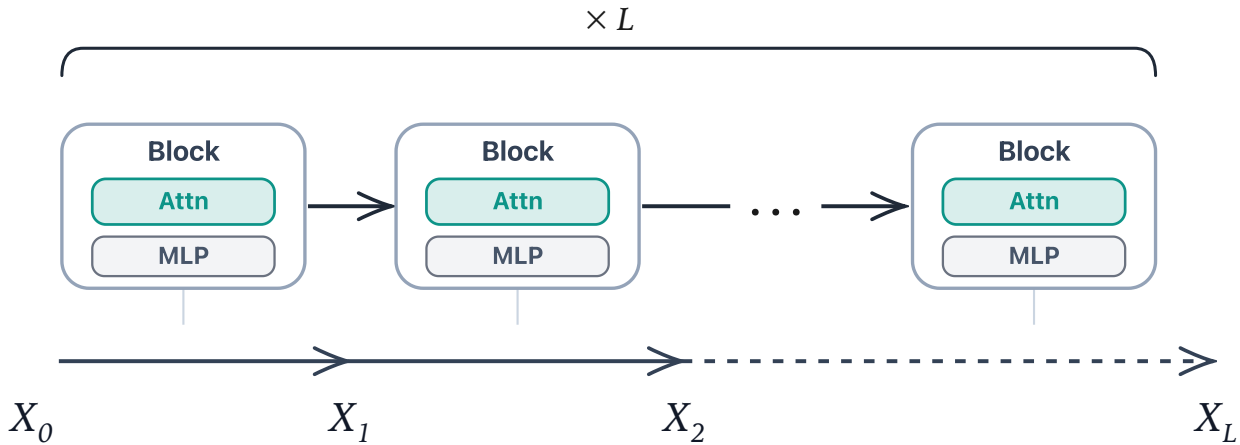
input tokens

$$x_1, x_2, \ldots, x_n \in \mathbb{R}^d$$

Repeat the Block

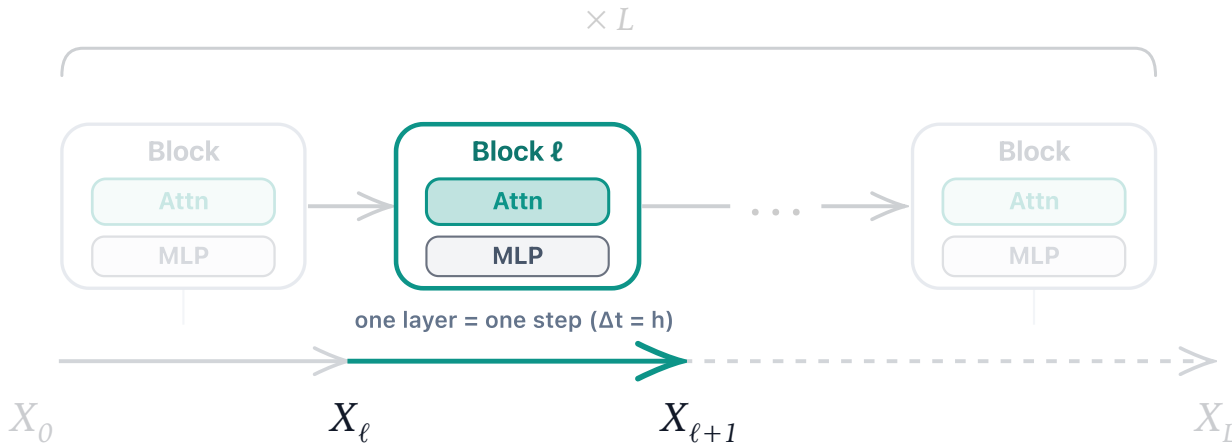


Repeat the Block



What happens across layers?

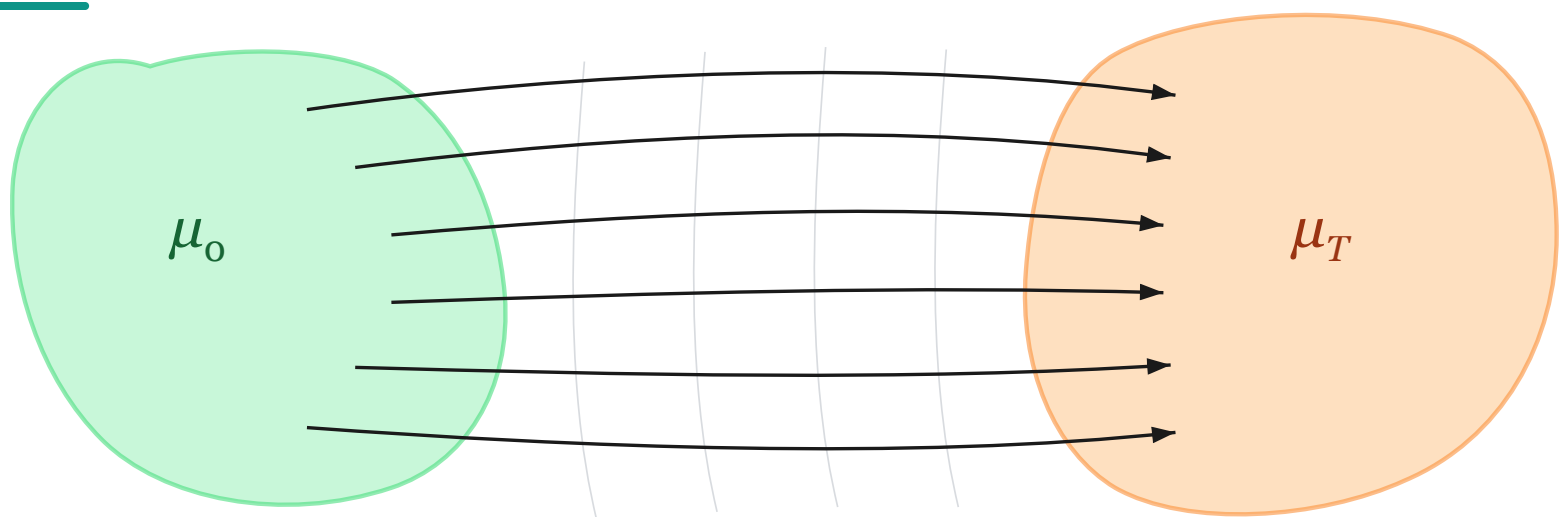
Repeat the Block



$$X_{\ell+1} = X_\ell + h\, F_\ell(X_\ell)$$

Transformer = Flow Map

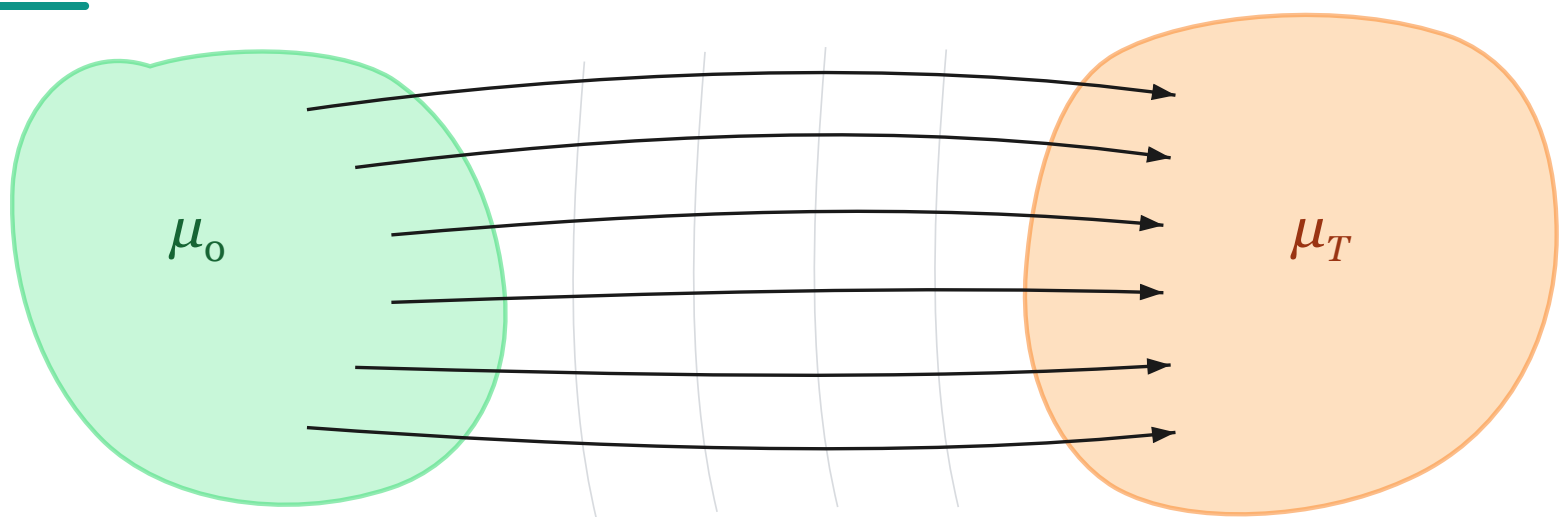
Transformer = Flow Map



Particle dynamics

$$\dot{x}_i(t) = X_t[\mu_t]\big(x_i(t)\big)$$

Transformer = Flow Map



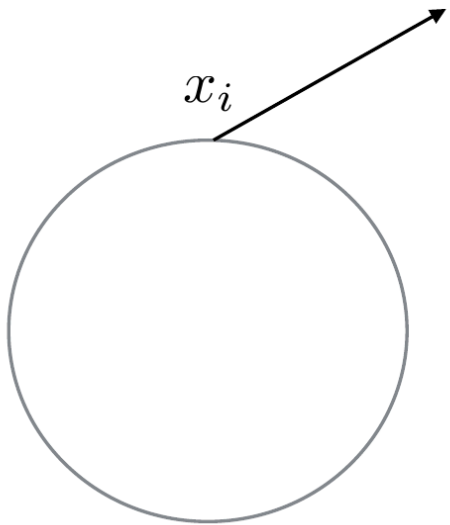
Particle dynamics

$$\dot{x}_i(t) = X_t[\mu_t]\big(x_i(t)\big)$$

Continuity equation

$$\partial_t \mu_t + \operatorname{div}\big(\mu_t\, X_t[\mu_t]\big) = 0$$

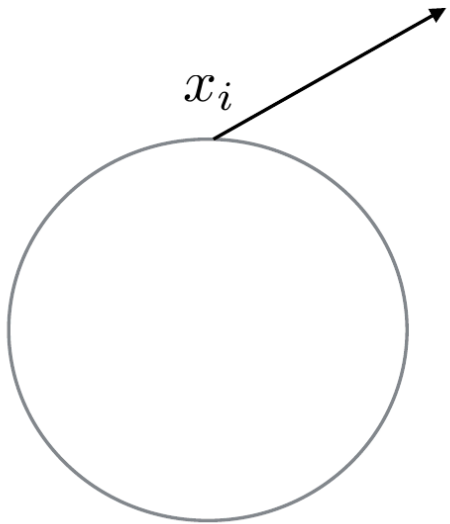
Self-Attention Dynamics



Transformer update

$$x_i \;\leftarrow\; x_i + \sum_{j=1}^{n} \frac{e^{\beta \langle Q_\ell x_i,\, K_\ell x_j \rangle}}{\sum_{k=1}^{n} e^{\beta \langle Q_\ell x_i,\, K_\ell x_k \rangle}}\, V_\ell\, x_j \;+\; \mathrm{MLP}$$

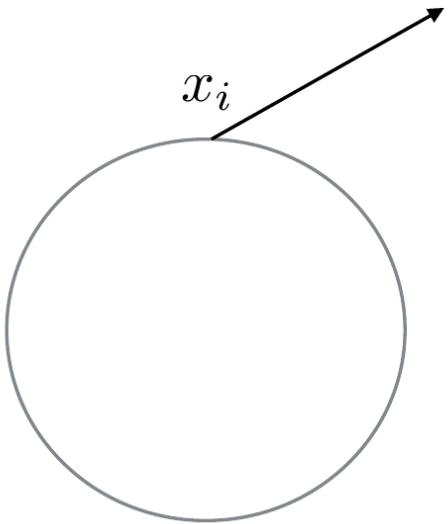
Self-Attention Dynamics



Transformer update

$$x_i \leftarrow x_i + \sum_{j=1}^{n} \frac{e^{\beta \langle Q_\ell x_i,\, K_\ell x_j \rangle}}{\sum_{k=1}^{n} e^{\beta \langle Q_\ell x_i,\, K_\ell x_k \rangle}} V_\ell\, x_j \;+\; \cancel{\text{MLP}}$$

Self-Attention Dynamics



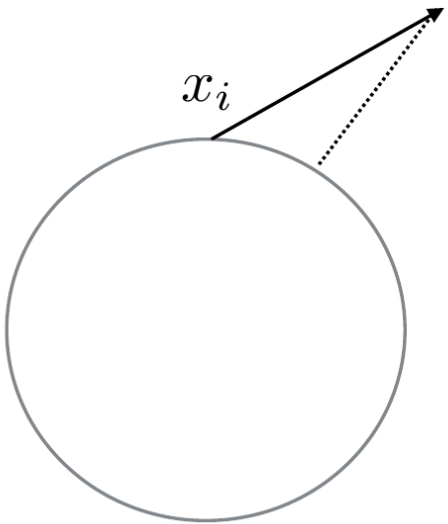
Transformer update

$$x_i \leftarrow x_i + \sum_{j=1}^{n} \frac{e^{\beta \langle \cancel{Q_\ell} x_i, \cancel{K_\ell} x_j \rangle}}{\sum_{k=1}^{n} e^{\beta \langle \cancel{Q_\ell} x_i, \cancel{K_\ell} x_k \rangle}} \cancel{V_\ell} x_j + \cancel{\text{MLP}}$$

Simplified step

$$x_i \leftarrow x_i + \sum_{j} \text{softmax}_j\big(\beta \, \langle x_i, \, x_j \rangle\big) \, x_j$$

Self-Attention Dynamics



LayerNorm: $\|x\| = 1$

Transformer update

$$x_i \leftarrow x_i + \sum_{j=1}^{n} \frac{e^{\beta \langle Q_\ell x_i, K_\ell x_j \rangle}}{\sum_{k=1}^{n} e^{\beta \langle Q_\ell x_i, K_\ell x_k \rangle}} V_\ell x_j + \text{MLP}$$

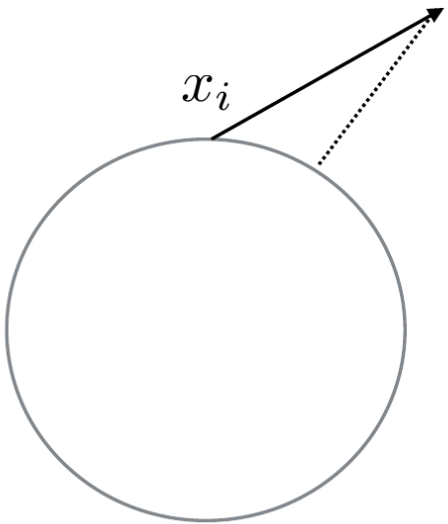
Simplified step

$$x_i \leftarrow x_i + \sum_{j} \text{softmax}_j\big(\beta \, \langle x_i, \, x_j \rangle\big) \, x_j$$

Gaussian distance kernel

$$\text{softmax}_j\big(\beta \, \langle x_i, \, x_j \rangle\big) \implies \text{softmax}_j\Big(-\tfrac{\beta}{2} \|x_i - x_j\|^2\Big)$$

Self-Attention Dynamics



LayerNorm: $\|x\| = 1$

Transformer update

$$x_i \leftarrow x_i + \sum_{j=1}^{n} \frac{e^{\beta \langle Q_\ell x_i, K_\ell x_j \rangle}}{\sum_{k=1}^{n} e^{\beta \langle Q_\ell x_i, K_\ell x_k \rangle}} V_\ell x_j + \text{MLP}$$

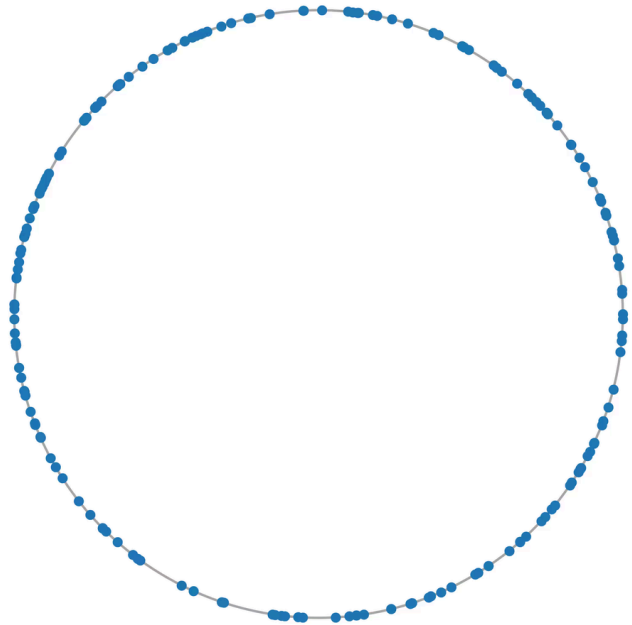
Simplified step

$$x_i \leftarrow x_i + \sum_{j} \text{softmax}_j\big(\beta \, \langle x_i, \, x_j \rangle\big) \, x_j$$

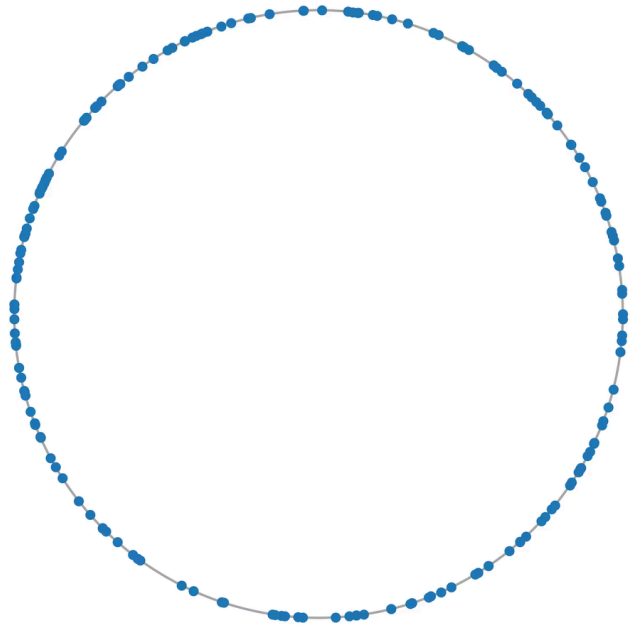
Gaussian distance kernel

$$\text{softmax}_j\big(\beta \, \langle x_i, \, x_j \rangle\big) \implies \text{softmax}_j\Big(-\tfrac{\beta}{2} \|x_i - x_j\|^2\Big)$$

Two Time-Scales

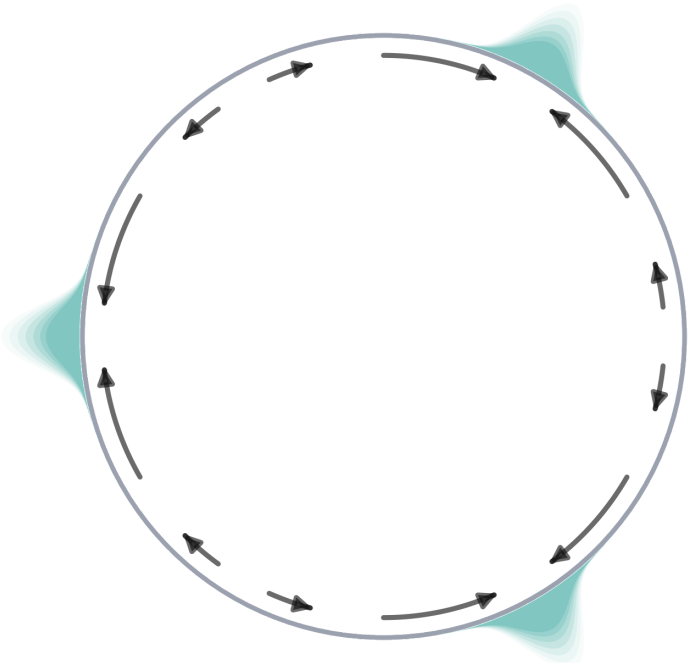


Two Time-Scales

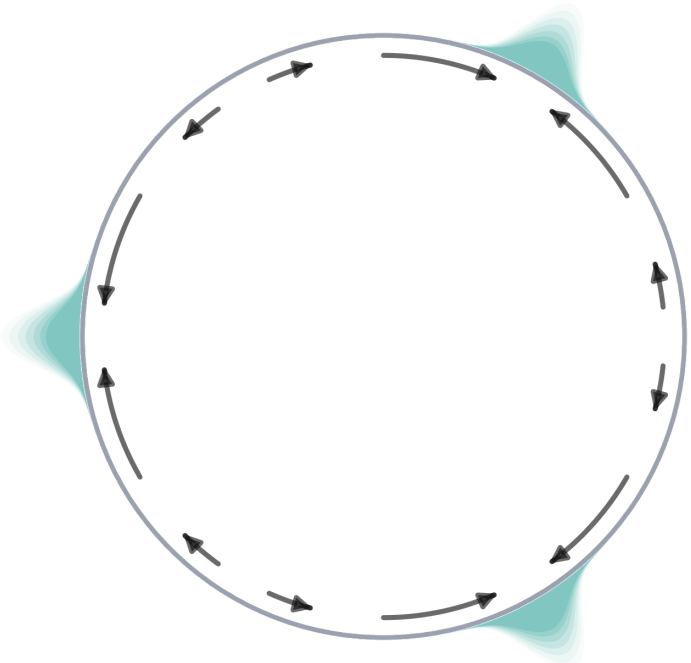


fast cluster formation → slow cluster-center merging

Softmax = Score Ascent



Softmax = Score Ascent



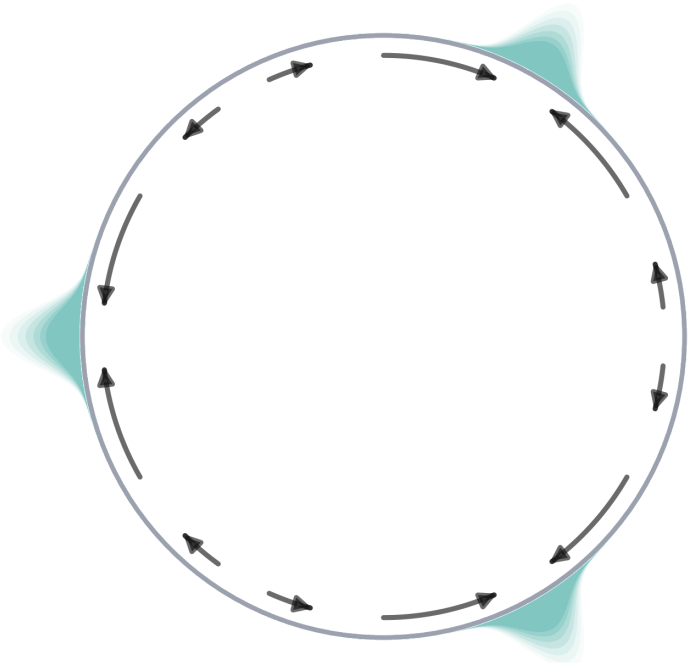
Gaussian blur (how much mass near x)

$$Z_\rho(x) \;=\; (K_\beta * \rho)(x), \qquad K_\beta(x,y) \;\propto\; e^{-\frac{\beta}{2}\|x-y\|^2}$$

Drift (direction of motion)

$$v_{\text{attn}}(x) \;=\; \tfrac{1}{\beta}\,\nabla \log\, Z_\rho(x)$$

Softmax = Score Ascent



Gaussian blur (how much mass near x)

$$Z_\rho(x) \;=\; (K_\beta * \rho)(x), \qquad K_\beta(x,y) \;\propto\; e^{-\frac{\beta}{2}\|x-y\|^2}$$

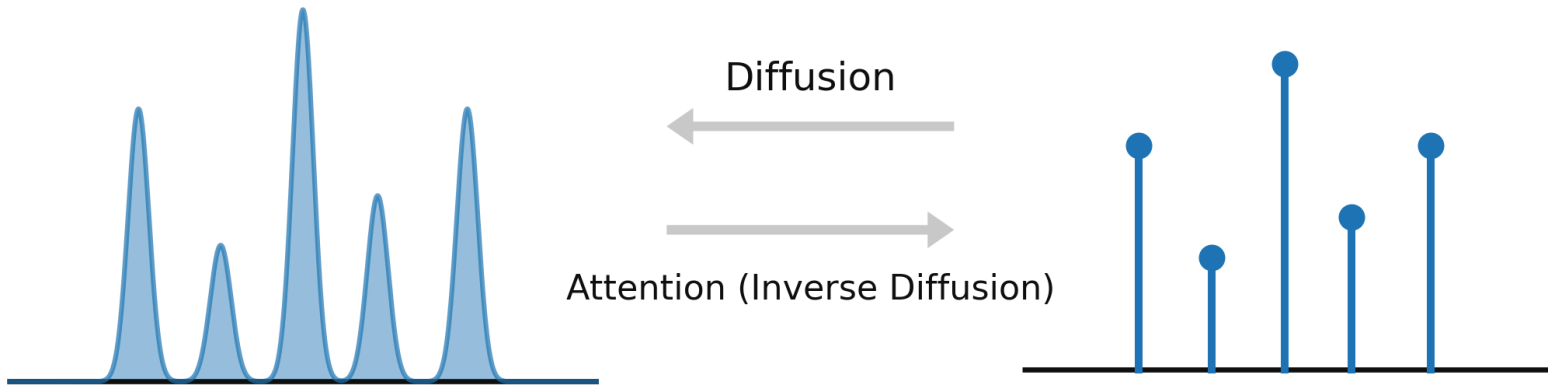
Drift (direction of motion)

$$v_{\text{attn}}(x) \;=\; \tfrac{1}{\beta}\,\nabla \log\, Z_\rho(x)$$

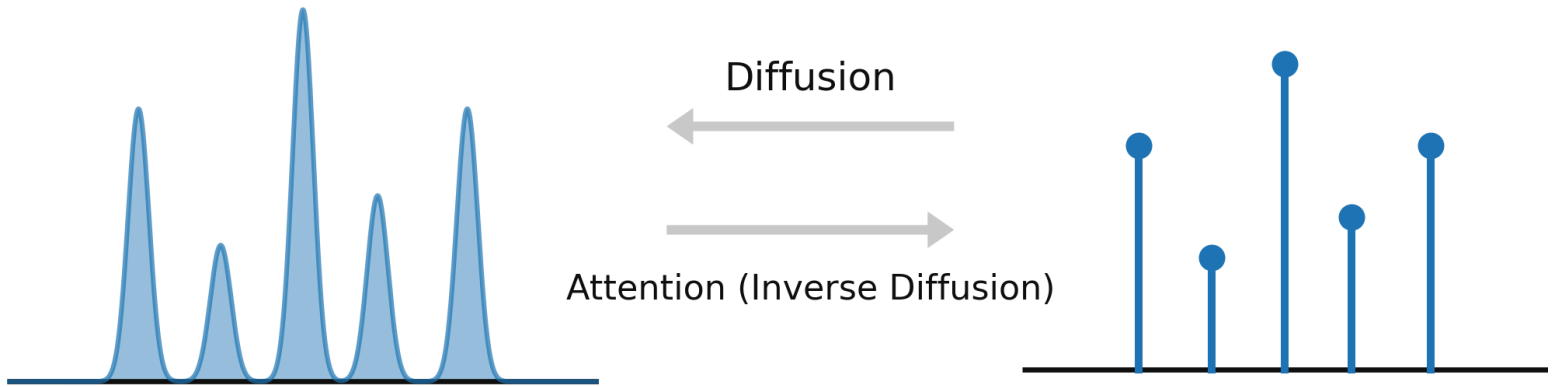
Heuristic (transport + narrow blur)

$$\begin{aligned} \partial_t \rho \;&=\; -\nabla\cdot\big(\rho\, v_{\text{attn}}\big) && \text{(continuity)} \\ &\approx\; -\tfrac{1}{\beta}\,\Delta\,\rho && (K_\beta * \rho \approx \rho) \end{aligned}$$

Attention Looks Like Inverse Heat



Attention Looks Like Inverse Heat



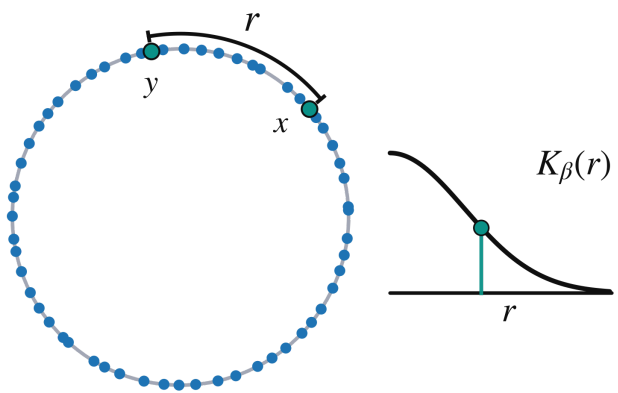
Diffusion

$$\partial_t \rho = +\Delta \rho$$

Inverse heat

$$\partial_t \rho = -\Delta \rho$$

Energy Goes Up

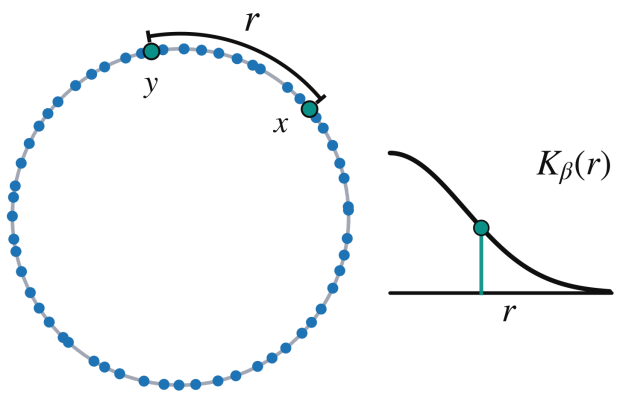


Pairwise energy

$$E_K[\mu] := \tfrac{1}{2} \iint K_\beta(x,y)\,\mu(dx)\,\mu(dy)$$

$$K_\beta(x,y) \propto e^{-\frac{\beta}{2}\|x-y\|^2}$$

Energy Goes Up



Pairwise energy

$$E_K[\mu] \;:=\; \tfrac{1}{2} \iint K_\beta(x,y)\,\mu(dx)\,\mu(dy)$$

$$K_\beta(x,y) \;\propto\; e^{-\frac{\beta}{2}\|x-y\|^2}$$

Unnormalized self-attention

$$\partial_t \mu_t \;\approx\; +\,\mathrm{grad}_{W_2}\, E_K[\mu_t]$$

$$\Longrightarrow \quad t \mapsto E_K[\mu_t] \text{ is increasing}$$

Entropy Goes Down

Energy ↑ (prev. slide) ⇒ Entropy ↓ (this slide).

Classical (Shannon) Entropy

$$H(\mu) := -\int \rho \log \rho$$

Theorem (Concentration Bookkeeping)

$$\frac{d}{dt} H(\mu_t) \le -(d-1)\Big(E_K[\mu_t] - E_K[\sigma_{\text{unif}}]\Big)$$

σ_{unif} = uniform measure on S^{d-1}

Entropy Goes Down

Energy ↑ (prev. slide) ⇒ Entropy ↓ (this slide).

Classical (Shannon) Entropy

$$H(\mu) := -\int \rho \log \rho$$

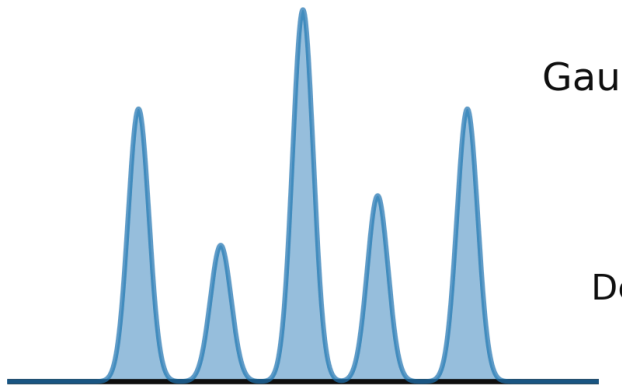
Theorem (Concentration Bookkeeping)

$$\frac{d}{dt} H(\mu_t) \le -(d-1)\Big(E_K[\mu_0] - E_K[\sigma_{\text{unif}}]\Big)$$

σ_{unif} = uniform measure on S^{d-1}

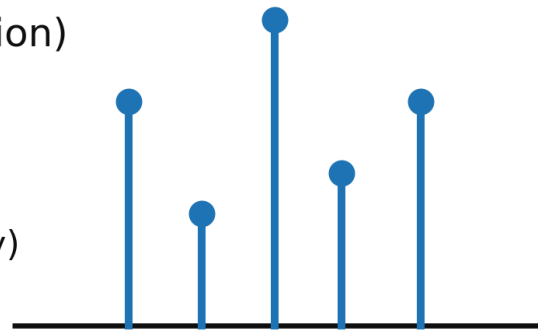
Corollary (linear decay). $H(\mu_t) \le H(\mu_0) - ct.$

Mixtures Are Blurred Atoms



Gaussian blur (convolution)

Deconvolution (recovery)



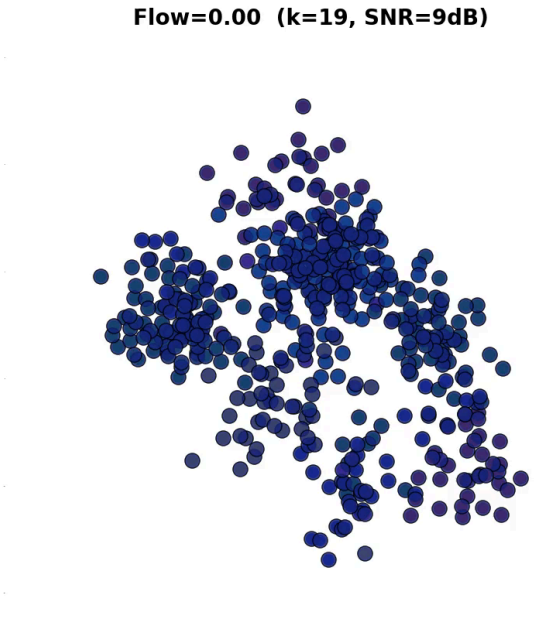
Observe

$\rho = \nu \star \varphi_\sigma$

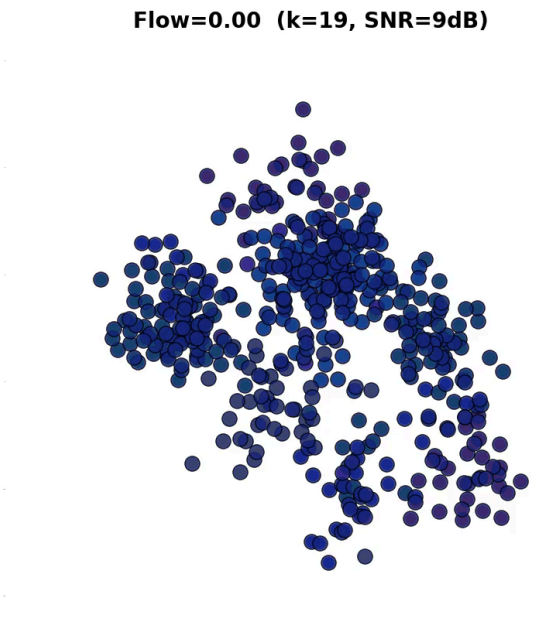
Recover

ν (atoms)

Flow speed controls concentration

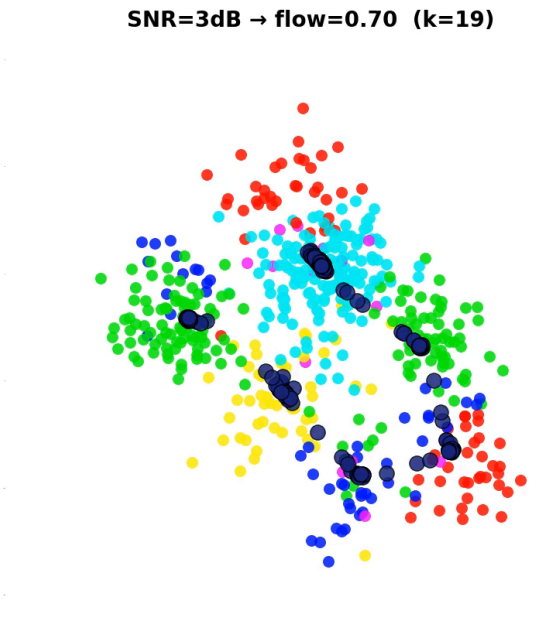


Flow speed controls concentration

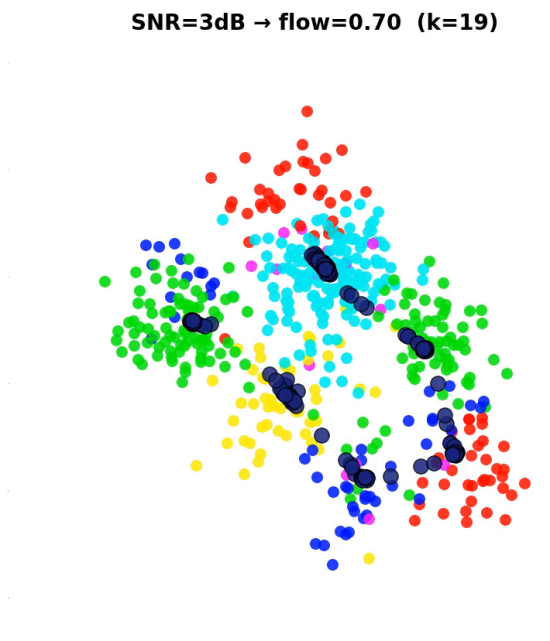


One trained model; `flow_speed` chooses how far we run the deblurring flow.

SNR controls flow speed



SNR controls flow speed

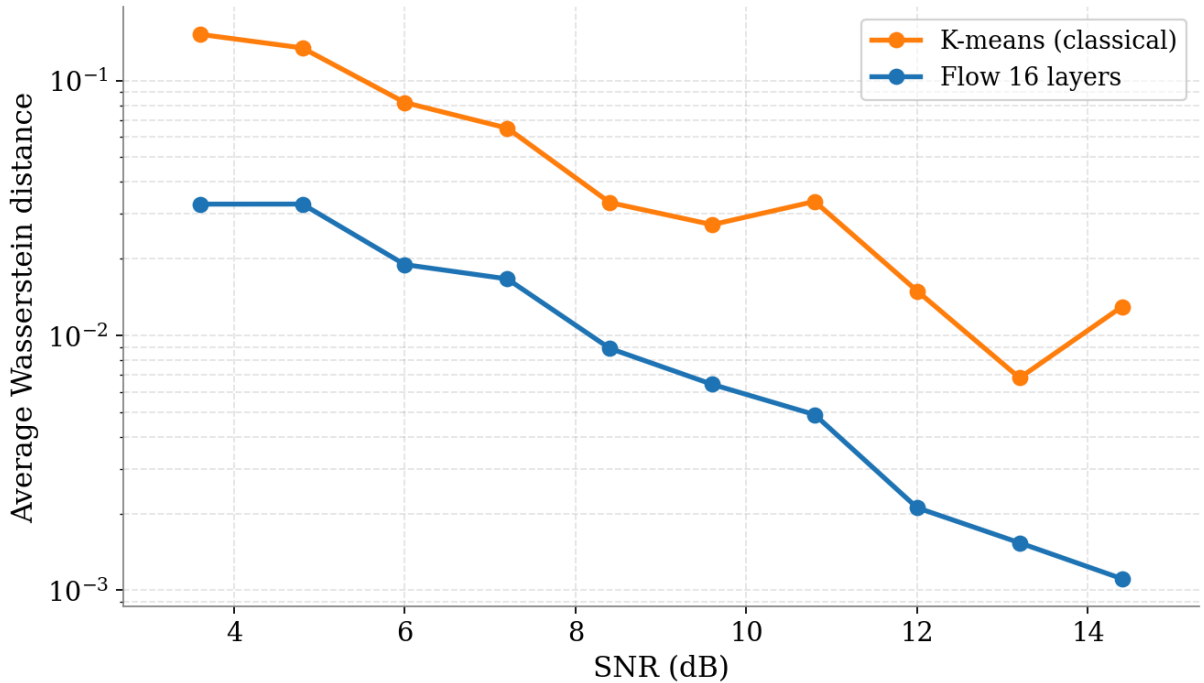


Input SNR → `flow_speed`

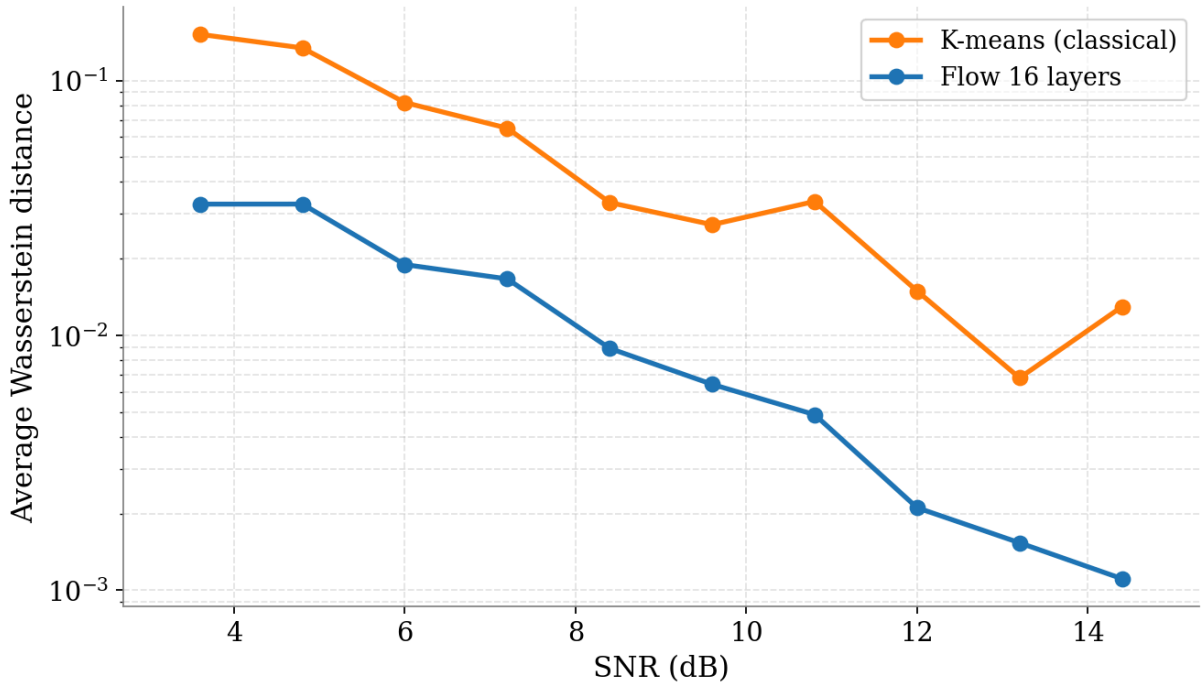
Low SNR → higher `flow_speed` → stronger clustering

High SNR → lower `flow_speed` → weaker clustering

Adaptive flow: extra clustering when needed

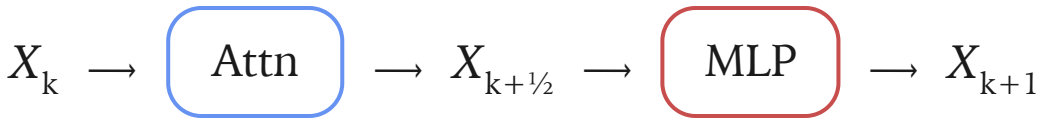


Adaptive flow: extra clustering when needed

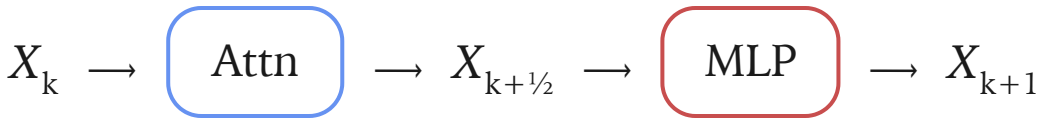


Largest improvement in the low-SNR regime.

Transformers as Optimizers

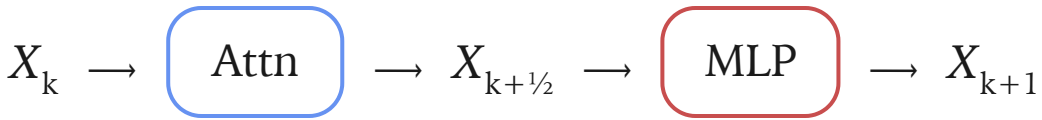


Transformers as Optimizers



$$
\begin{aligned}
X_{k+\frac{1}{2}} &= X_k + \eta \cdot \mathrm{Attn}(X_k) \\
X_{k+1} &= X_{k+\frac{1}{2}} + \eta \cdot \mathrm{MLP}(X_{k+\frac{1}{2}})
\end{aligned}
$$

Transformers as Optimizers



$$X_{k+\frac{1}{2}} = X_k + \eta \cdot \mathrm{Attn}(X_k)$$
$$X_{k+1} = X_{k+\frac{1}{2}} + \eta \cdot \mathrm{MLP}(X_{k+\frac{1}{2}})$$

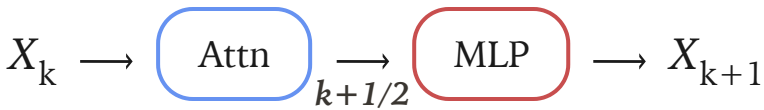
gradient ascent on $E = E_{Attn} + E_{MLP}$

$\mathrm{Attn} \approx \nabla E_{Attn}$ — interaction

$\mathrm{MLP} \approx \nabla E_{MLP}$ — per-token

Discretization = Architecture

LIE–TROTTER (SEQUENTIAL)

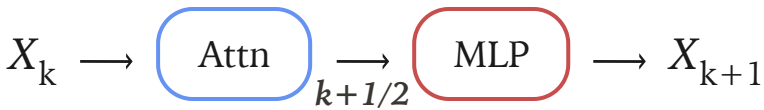


EULER (COMBINED)

PaLM, early Phis

Discretization = Architecture

LIE–TROTTER (SEQUENTIAL)



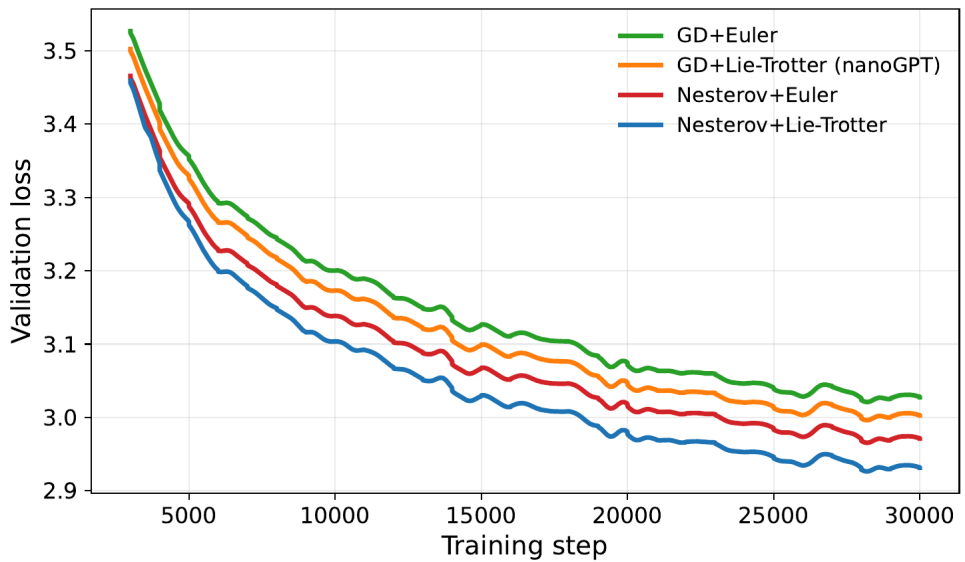
EULER (COMBINED)

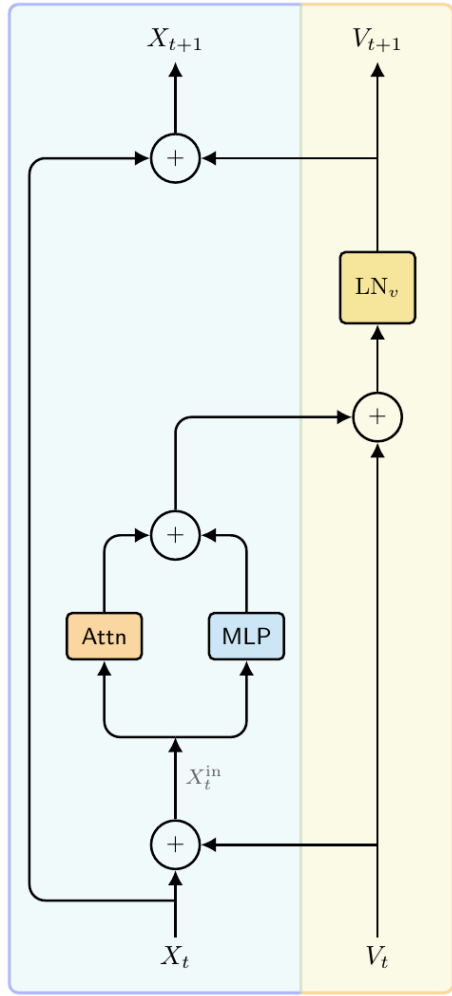
PaLM, early Phis

Now swap GD → momentum (Nesterov)

YuriiFormer

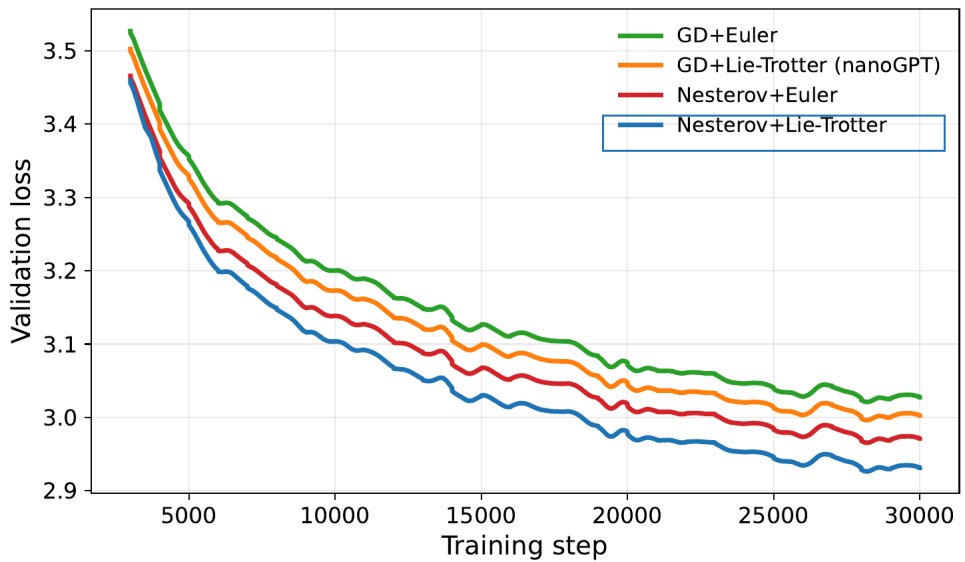
Nesterov + Lie-Trotter on $E_{\text{Attn}} + E_{\text{MLP}}$



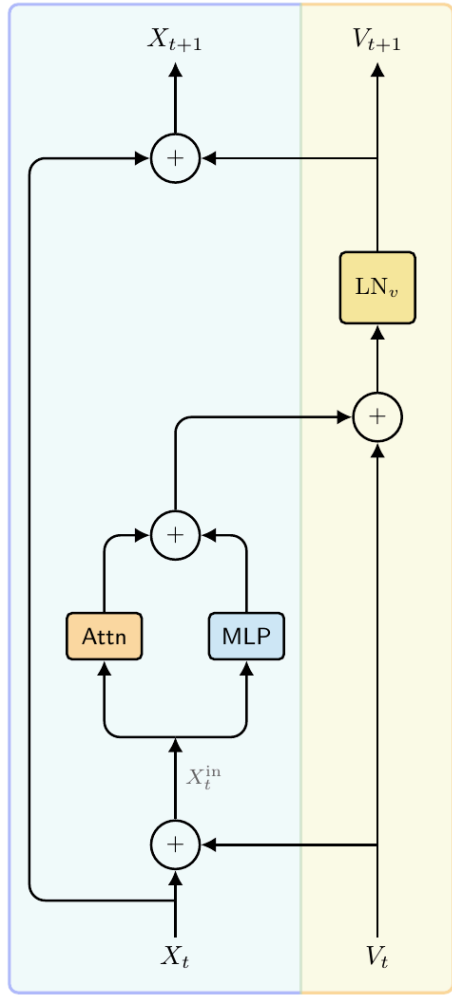


YuriiFormer

Nesterov + Lie-Trotter on $E_{\text{Attn}} + E_{\text{MLP}}$



Lowest validation loss (matched compute, same params)



CONCLUSION

AI for Mathematics

Until now: math → AI. Now: AI → math.

A Real Ask

COLLABORATOR

Attached: arXiv:2512.02412. I think the gap between n^5 and n^6 could be easy to close. Could AI help?

A Real Ask

COLLABORATOR

Attached: arXiv:2512.02412. I think the gap between n^5 and n^6 could be easy to close. Could AI help?

Input

message + paper

→

Human Control

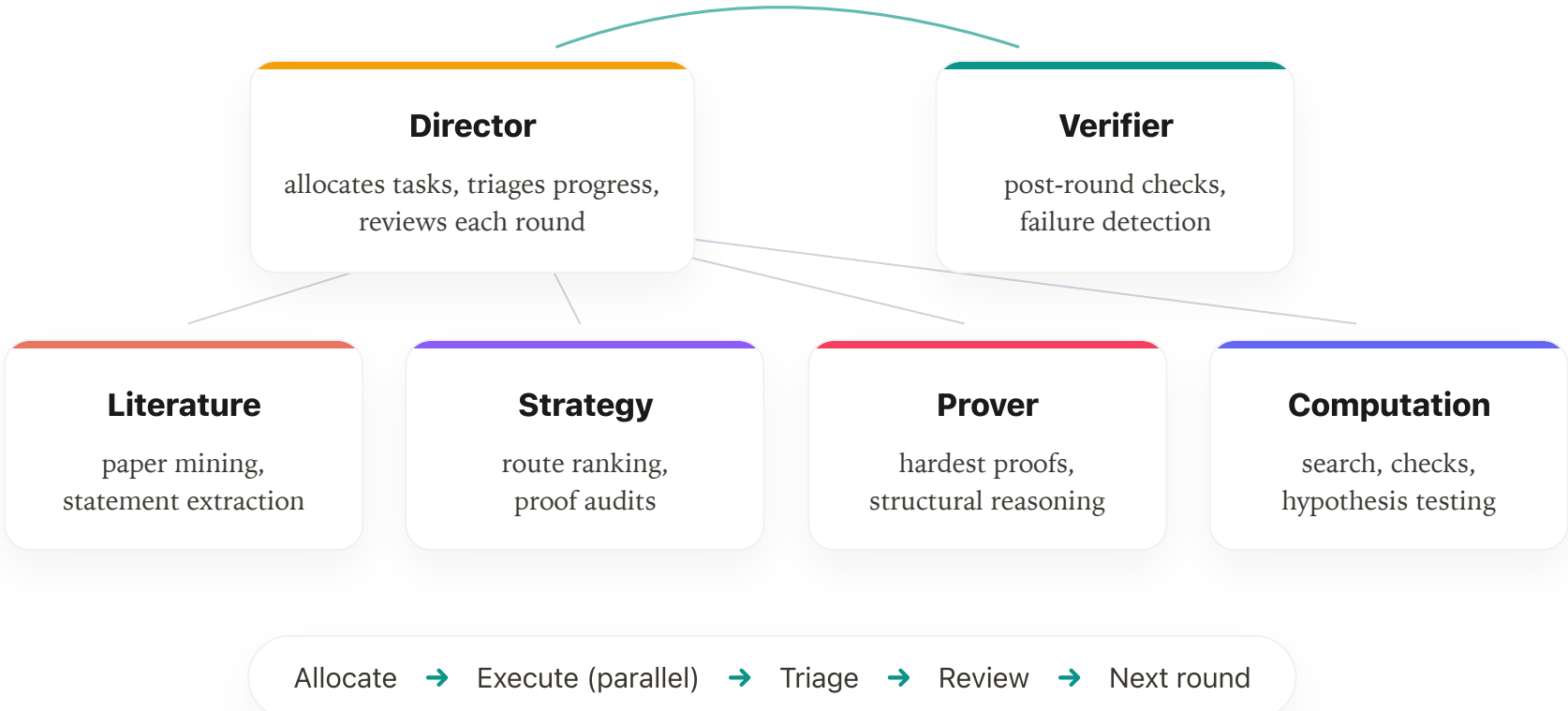
I write the blueprint. Agents execute in parallel. Verification keeps the loop honest.

→

Output

draft + checks

Make the Loop Explicit



Why Split Agents?

Director

Dispatch independent tasks in parallel; use computation to kill bad routes early.

Why Split Agents?

Director

Dispatch independent tasks in parallel; use computation to kill bad routes early.

dispatches to 4 workers in parallel

Literature

Find the right theorem.

→ Statement + exact assumptions.

Strategy

Generate approaches to test.

→ Ranked route list.

Prover

Close the gap.

→ Lemma + proof.

Computation

COUNTEREXAMPLE

Try to break the route.

→ Counterexample to the suggested approach (or evidence it survives).

Verification in Action

Write

Worker writes a proof of the equivalence theorem.

Verify → Fail

Verifier finds an invalid inference.

```
[L-Equiv]: FAIL
invalid inference: "c_i ∈ [n] ⇒ x_i = c_i"
for latent-label witnesses
```

Verification in Action

Write

Worker writes a proof of the equivalence theorem.

Verify → Fail

Verifier finds an invalid inference.

```
[L-Equiv]: FAIL
invalid inference: "c_i ∈ [n] ⇒ x_i = c_i"
for latent-label witnesses
```

Diagnose & Repair

All four workers converge on the same fix: strengthen the invariant.

Verify → Pass

Re-verified. Theorem complete.

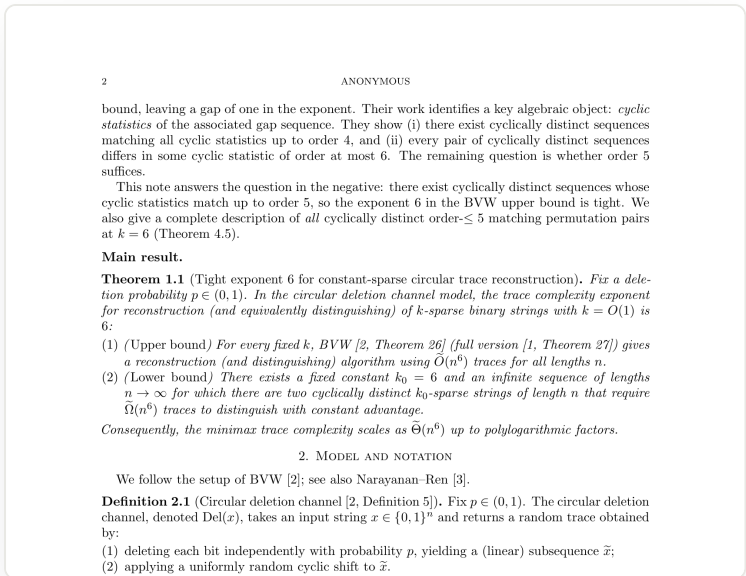
Self-correcting system. Write → verify → fail → diagnose → repair → verify → pass.

Endgame

Lean kernel: **ACCEPT** / **REJECT**

agents write candidates; the kernel checks.

Result: The Paper



2 ANONYMOUS

bound, leaving a gap of one in the exponent. Their work identifies a key algebraic object: *cyclic statistics* of the associated gap sequence. They show (i) there exist cyclically distinct sequences matching all cyclic statistics up to order 4, and (ii) every pair of cyclically distinct sequences differs in some cyclic statistic of order at most 6. The remaining question is whether order 5 suffices.

This note answers the question in the negative: there exist cyclically distinct sequences whose cyclic statistics match up to order 5, so the exponent 6 in the BVW upper bound is tight. We also give a complete description of *all* cyclically distinct order-≤ 5 matching permutation pairs at $k = 6$ (Theorem 4.5).

Main result.

Theorem 1.1 (Tight exponent 6 for constant-sparse circular trace reconstruction)**.** *Fix a deletion probability $p \in (0, 1)$. In the circular deletion channel model, the trace complexity exponent for reconstruction (and equivalently distinguishing) of k-sparse binary strings with $k = O(1)$ is 6:*

(1) *(*Upper bound*) For every fixed k, BVW [2, Theorem 26] (full version [1, Theorem 27]) gives a reconstruction (and distinguishing) algorithm using $\widetilde{O}(n^6)$ traces for all lengths n.*
(2) *(*Lower bound*) There exists a fixed constant $k_0 = 6$ and an infinite sequence of lengths $n \to \infty$ for which there are two cyclically distinct k_0-sparse strings of length n that require $\widetilde{\Omega}(n^6)$ traces to distinguish with constant advantage.*

Consequently, the minimax trace complexity scales as $\widetilde{\Theta}(n^6)$ up to polylogarithmic factors.

2. Model and notation

We follow the setup of BVW [2]; see also Narayanan–Ren [3].

Definition 2.1 (Circular deletion channel [2, Definition 5])**.** Fix $p \in (0, 1)$. The circular deletion channel, denoted $\mathrm{Del}(x)$, takes an input string $x \in \{0, 1\}^n$ and returns a random trace obtained by:

(1) deleting each bit independently with probability p, yielding a (linear) subsequence $\widetilde{x}$;
(2) applying a uniformly random cyclic shift to $\widetilde{x}$.

2 rounds, 3 hours

COLLABORATOR VERDICT

Everything is great, except: you can **remove the permutation condition**.

Any pair of sextuplets with equal orders up to five gives an example.

Artifact is correct. Remaining work is insight.

insight → rerun loop → stronger theorem

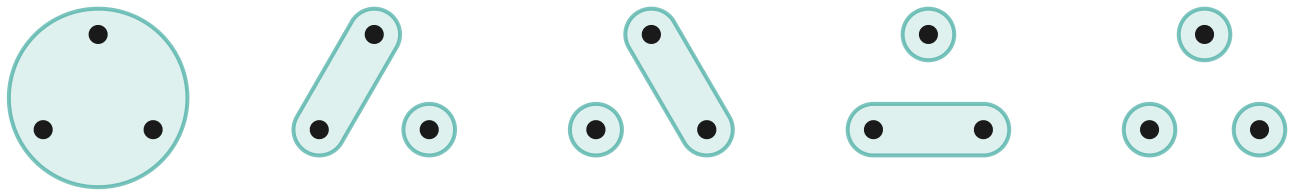
Thank you

BACKUP

Backup

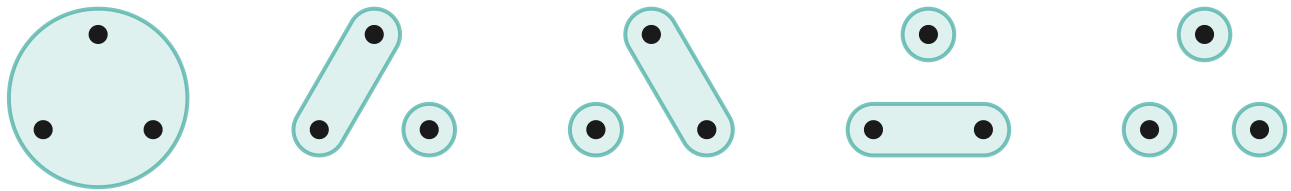
What we can observe

A random configuration induces a random partition of the terminals.



What we can observe

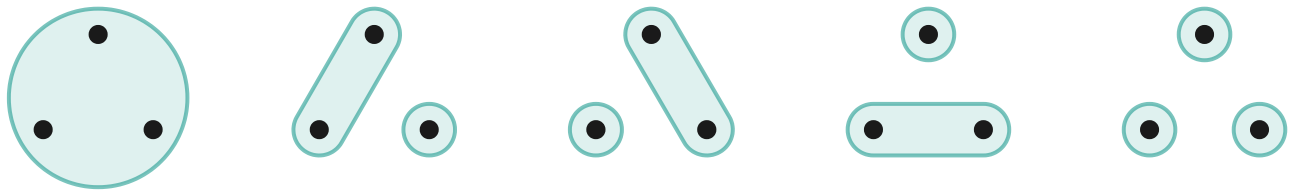
A random configuration induces a random partition of the terminals.



abc $ab \mid c$ $ac \mid b$ $a \mid bc$ $a \mid b \mid c$

What we can observe

A random configuration induces a random partition of the terminals.



abc $\quad ab \mid c$ $\quad ac \mid b$ $\quad a \mid bc$ $\quad a \mid b \mid c$

Let Π be the terminal partition. $\quad \mu(\pi) := \mathbb{P}(\Pi = \pi)$.

Which μ can bond percolation achieve?

Classical inequalities for bond percolation

Can bonds simulate a 4-hyperedge?

$abcd$ $\qquad$ $a \mid b \mid c \mid d$

Classical inequalities for bond percolation

FKG. $\mathbb{P}(A \cap B) \geq \mathbb{P}(A)\,\mathbb{P}(B)$ *(increasing A, B)*

Can bonds simulate a 4-hyperedge?

$abcd$ $\qquad$ $a \mid b \mid c \mid d$

Classical inequalities for bond percolation

FKG. $\mathbb{P}(A \cap B) \geq \mathbb{P}(A)\,\mathbb{P}(B)$ *(increasing A, B)*

BK / Reimer. $\mathbb{P}(A \,\square\, B) \leq \mathbb{P}(A)\,\mathbb{P}(B)$

($\square$ = disjoint occurrence; BK: incr.; Reimer: all)

Can bonds simulate a 4-hyperedge?

$abcd$ $\qquad$ $a \mid b \mid c \mid d$

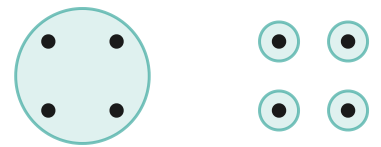
Classical inequalities for bond percolation

FKG. $\mathbb{P}(A \cap B) \geq \mathbb{P}(A)\,\mathbb{P}(B)$ *(increasing A, B)*

BK / Reimer. $\mathbb{P}(A \,\square\, B) \leq \mathbb{P}(A)\,\mathbb{P}(B)$

($\square$ = disjoint occurrence; BK: incr.; Reimer: all)

Can bonds simulate a 4-hyperedge?



$abcd$ $\qquad$ $a \mid b \mid c \mid d$

Proposition (4-hyperedge obstruction)

$$\mu(abcd) \;\leq\; \big(1 - \mu(a \mid b \mid c \mid d)\big)^2$$

Strong enough to rule out 4-hyperedge simulation — but not 3.

A 3-hyperedge obstruction

abc

$a \mid b \mid c$

A 3-hyperedge obstruction

abc $\qquad\qquad$ $a \mid b \mid c$

Theorem (Gladkov–Zimin, 2026)

For bond percolation on a graph G with terminals a, b, c:

$$\mu(abc)\,\mu(a \mid b \mid c) \;\leq\; 1 - \mu(abc) - \mu(a \mid b \mid c)$$

In particular: $\mu(abc) = p$, $\mu(a \mid b \mid c) = 1 - p$, $0 < p < 1$ is impossible.

A 3-hyperedge obstruction

abc $\qquad\qquad$ $a \mid b \mid c$

Theorem (Gladkov–Zimin, 2026)

For bond percolation on a graph G with terminals a, b, c:

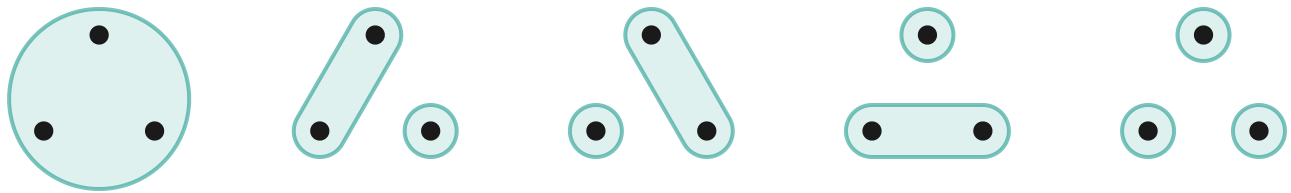
$$\mu(abc)\,\mu(a \mid b \mid c) \;\leq\; 1 - \mu(abc) - \mu(a \mid b \mid c)$$

In particular: $\mu(abc) = p$, $\mu(a \mid b \mid c) = 1 - p$, $0 < p < 1$ is impossible.

Exact simulation fails — Hollom's construction only needs a weaker gadget condition.

Computer-assisted inequalities

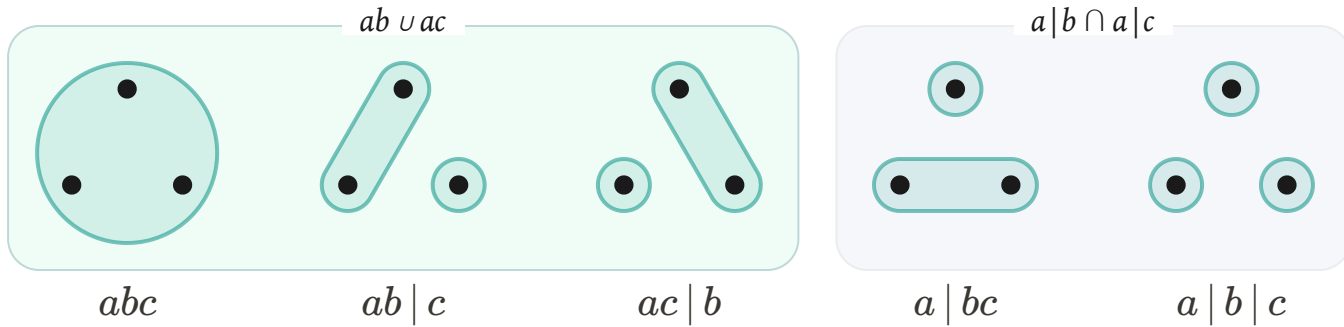
Oracle + double description enumerates facet inequalities for μ.



abc $ab \mid c$ $ac \mid b$ $a \mid bc$ $a \mid b \mid c$

Computer-assisted inequalities

Oracle + double description enumerates facet inequalities for μ.

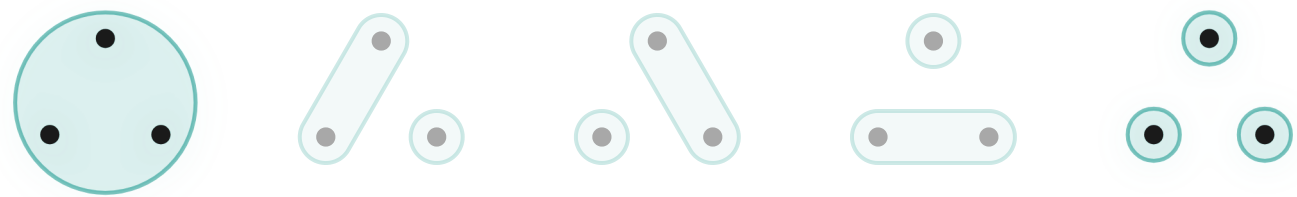


Proposition (Computer-discovered strengthening of the 3-terminal obstruction)

$$\mu(a\,|\,b\cap a\,|\,c)\,\mu(ab\cup ac)\;\le\;\mu(ab|c)+\mu(ac|b)+\mu(a|bc)-\mu(ab|c)^2-\mu(ac|b)^2.$$

Computer-assisted inequalities

Oracle + double description enumerates facet inequalities for μ.

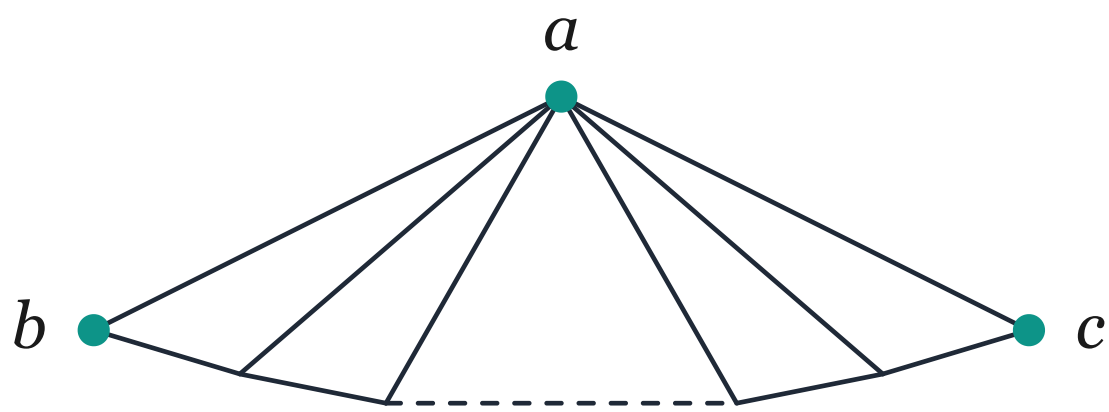


Theorem (Aas conjecture, proved by Gladkov–Zimin)

Stronger version of FKG for connectivity events.

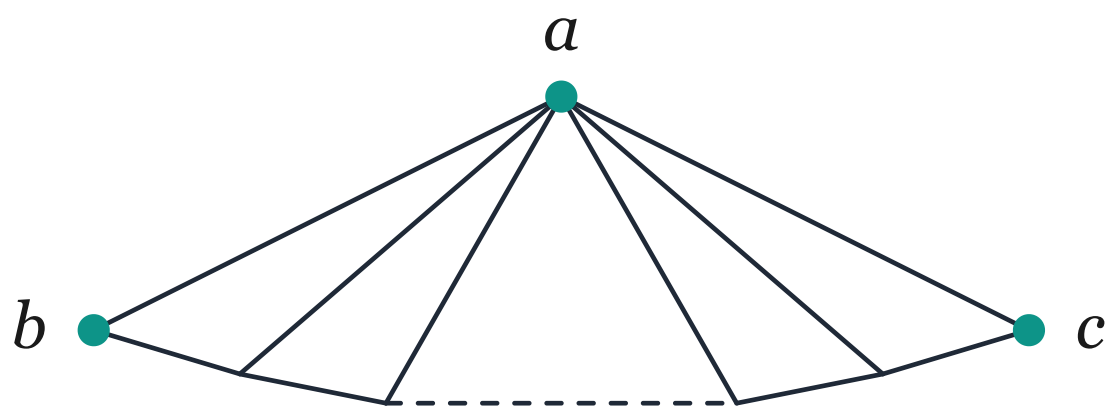
$$\mu(abc)\,\mu(a \mid b \mid c) \;\geq\; \mu(ab|c)\mu(ac|b) + \mu(ab|c)\mu(a|bc) + \mu(ac|b)\mu(a|bc).$$

Back to bunkbed: the gadget



Planar gadget G_n with terminals a, b, c.

Back to bunkbed: the gadget



Planar gadget G_n with terminals a, b, c.

Lemma (Covariance amplification gadget)

$$\mathrm{Cov}\big(a \leftrightarrow b,\ a \leftrightarrow c\big) \ > \ \left(\frac{n}{3} - 1\right) \mu(a \,|\, bc).$$