The emergence of clusters in self-attention dynamics

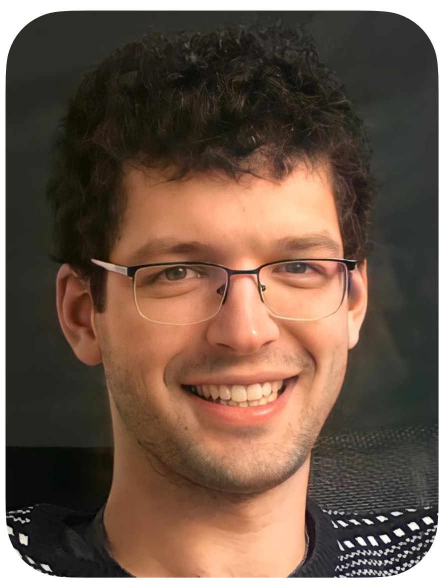
arXiv:2305.05465

Borjan[1] Geshkovski
MIT

Workshop EDP-COSy, LAAS
Oct 20, 2023

[1] BorYann

Based on joint work with



Cyril Letrouit
(CNRS, Orsay)

Yury Polyanskiy
(MIT)

Philippe Rigollet
(MIT)

Machine learning

Approximate **unknown** function $f : \mathbb{R}^d \to \mathbb{R}^m$.

Have access to **data**

$$\left\{x^{(i)}, f(x^{(i)})\right\}_{i\in[N]} \subset \mathbb{R}^d \times \mathbb{R}^m$$

Propose an **architecture** $f_\theta : \mathbb{R}^d \to \mathbb{R}^m$ depending on **parameters** $\theta \in \mathbb{R}^p$ and solve

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \left\| f_\theta(x^{(i)}) - f(x^{(i)}) \right\|^2 + \|\theta\|_{\text{some norm}}$$

Machine learning

Approximate **unknown** function $f : \mathbb{R}^d \to \mathbb{R}^m$.

Have access to **data**

$$\left\{x^{(i)}, f(x^{(i)})\right\}_{i\in[N]} \subset \mathbb{R}^d \times \mathbb{R}^m$$

Propose an **architecture** $f_\theta : \mathbb{R}^d \to \mathbb{R}^m$ depending on **parameters** $\theta \in \mathbb{R}^p$ and solve

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \left\| f_\theta(x^{(i)}) - f(x^{(i)}) \right\|^2 + \|\theta\|_{\text{some norm}}$$

Stats: $x^{(i)}$ random → "law of large numbers" → something

Optim: algorithm for finding θ^* → something

Control: ???

Neural networks

1. **Feed-forward networks**: for any $i \in [N]$,

$$\begin{cases} x_i(t+1) = c(t)\sigma\Big(a(t)x_i(t) + b(t)\Big) & t \geqslant 1 \text{ integer} \\ x_i(0) = x^{(i)} \end{cases}$$

   - $x_i(t) \in \mathbb{R}^{d_t}$
   - $a(t) \in \mathbb{R}^{d_t \times d_t}$, $b(t) \in \mathbb{R}^{d_t}$, $c(t) \in \mathbb{R}^{d_{t+1} \times d_t}$
   - $\sigma \in C^{0,1}(\mathbb{R})$ element-wise ($\sigma(x) = (x)_+$ or $\sigma(x) = \tanh(x)$)

So $\theta = (a(t), b(t), c(t))_{t \geqslant 1}$ and

$$\boxed{f_\theta(x^{(i)}) = \mathbf{P}x_i(T)}$$

for some projector $\mathbf{P} : \mathbb{R}^{d_T} \to \mathbb{R}^m$

Neural networks

1. **Feed-forward networks**: for any $i \in [N]$,

$$\begin{cases} x_i(t+1) = c(t)\sigma\Big(a(t)x_i(t) + b(t)\Big) & t \geqslant 1 \text{ integer} \\ x_i(0) = x^{(i)} \end{cases}$$

   - $x_i(t) \in \mathbb{R}^{d_t}$
   - $a(t) \in \mathbb{R}^{d_t \times d_t}$, $b(t) \in \mathbb{R}^{d_t}$, $c(t) \in \mathbb{R}^{d_{t+1} \times d_t}$
   - $\sigma \in C^{0,1}(\mathbb{R})$ element-wise ($\sigma(x) = (x)_+$ or $\sigma(x) = \tanh(x)$)

   So $\theta = (a(t), b(t), c(t))_{t \geqslant 1}$ and

$$\boxed{f_\theta(x^{(i)}) = \mathbf{P}x_i(T)}$$

   for some projector $\mathbf{P} : \mathbb{R}^{d_T} \to \mathbb{R}^m$

2. **Residual networks**:

$$x_i(t+1) = x_i(t) + c(t)\sigma\Big(a(t)x_i(t) + b(t)\Big)$$

Neural ODEs

Idealize to continuous time ([E '17], [Sontag-Sussmann '97]):

$$\begin{cases} \dot{x}_i(t) = c(t)\sigma(a(t)x_i(t) + b(t)) & t \in [0, T] \\ x_i(0) = x^{(i)} \end{cases}$$

ML $\iff$ **control** of **many** initial conditions $x^{(i)}$ to targets $f(x^{(i)})$ by a **single control** (a, b, c)

Neural ODEs

Idealize to continuous time ([E '17], [Sontag-Sussmann '97]):

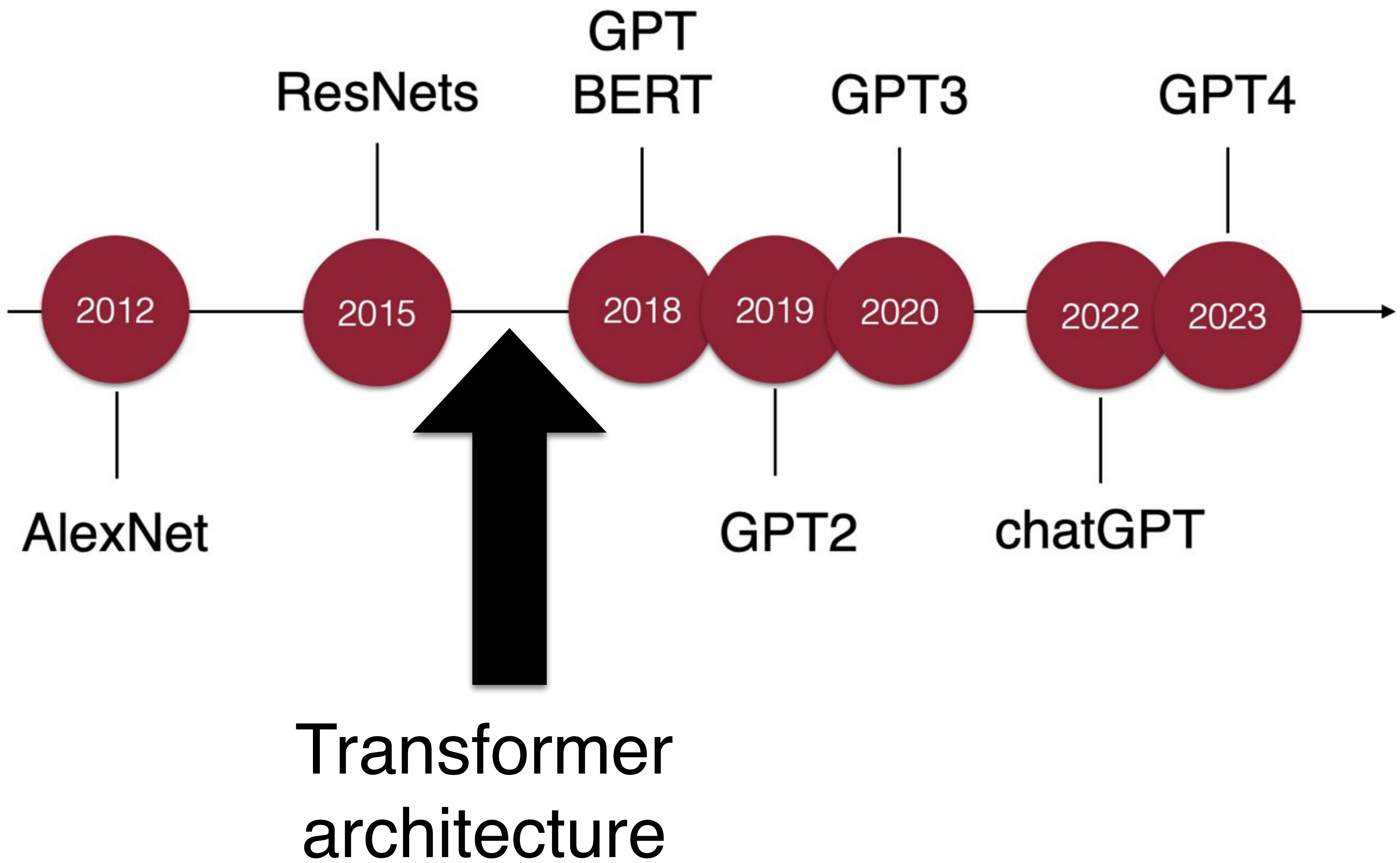
$$\begin{cases} \dot{x}_i(t) = c(t)\sigma(a(t)x_i(t) + b(t)) & t \in [0, T] \\ x_i(0) = x^{(i)} \end{cases}$$

ML $\iff$ **control** of **many** initial conditions $x^{(i)}$ to targets $f(x^{(i)})$ by a **single control** (a, b, c)

- [Ruiz-Balet, Zuazua '23], [Li, Lin, Shen '22]: exact controllability of any N initial points to any N targets, in any time $T > 0$
- [Ruiz-Balet, Zuazua '23], [Li, Lin, Shen '22]: for any $\varepsilon > 0$, $f \in L^2(\mathbb{R}^d; \mathbb{R}^m)$, there exist bounded controls $\theta = (a, b, c)$ s.t. $\|f - \Phi_\theta^t\|_{L^2} \leqslant \varepsilon$
- [G. '21] Optimal control: rates of error in terms of T
- [Agrachev, Sarychev '22], [Scagliotti '22] partial Lie brackets results
- and (not many) others (Bonnet, Cipriani, ...)

Light years away from a systematic theory and sharp results

The Transformer



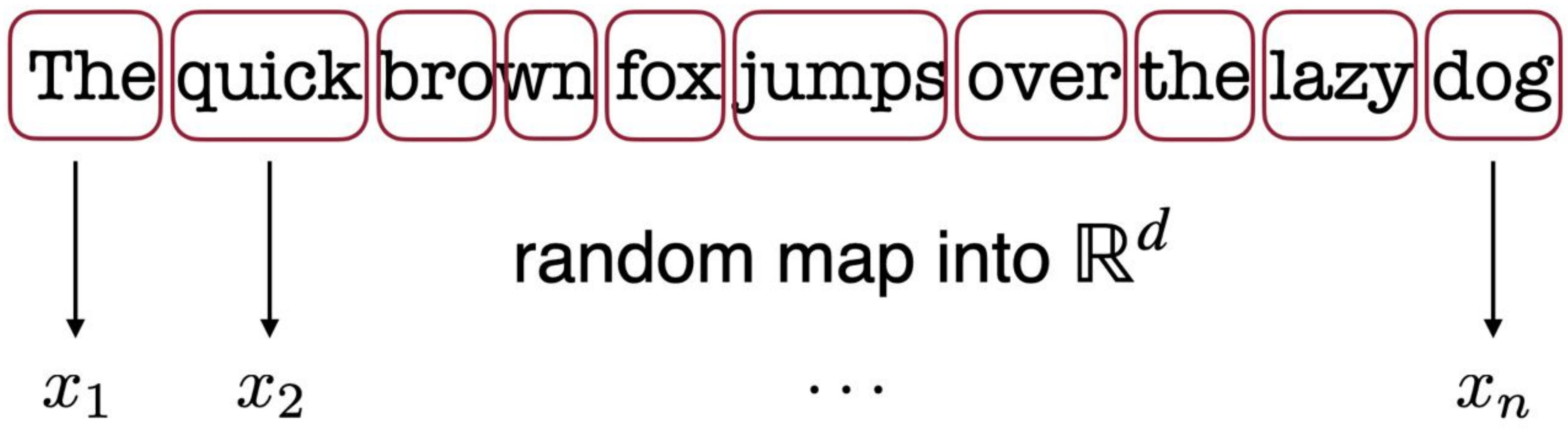
Attention is **all you need** [PDF] neurips.cc

A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc

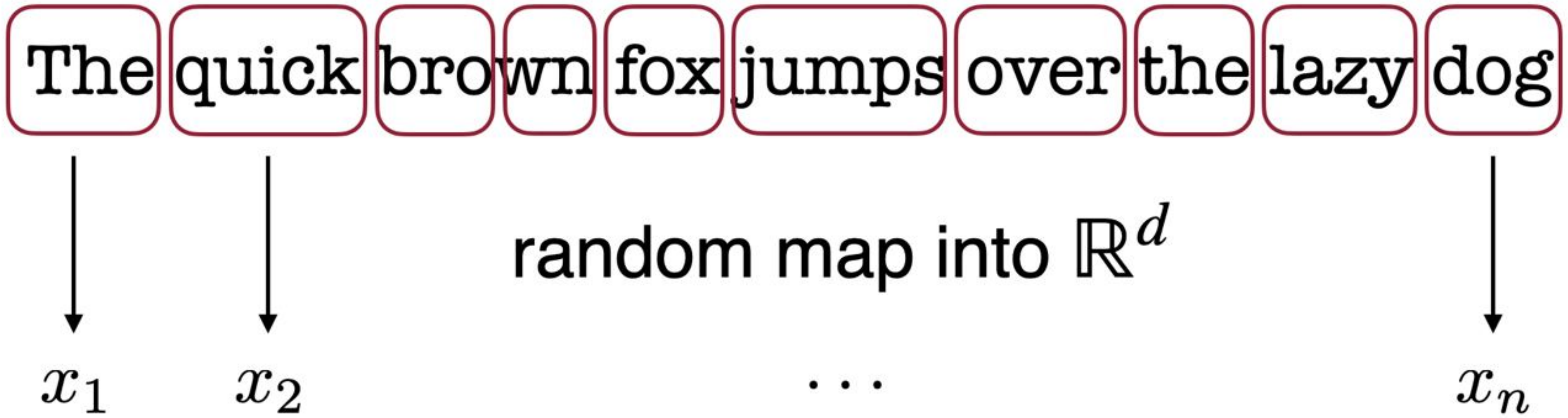
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent … **We** implement this inside of scaled dot-product **attention** by masking out (setting to $-\infty$) …

☆ Save 99 Cite Cited by 93154 Related articles All 62 versions »

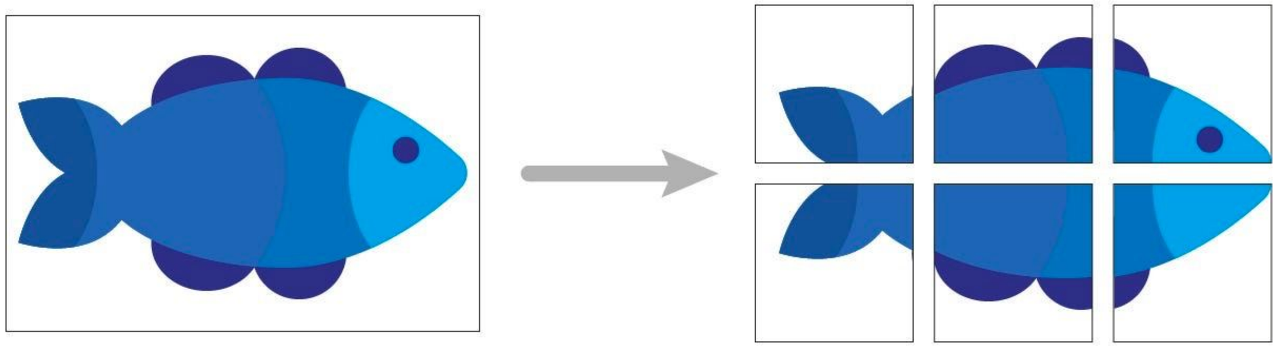
The Transformer: tokens and prompts



The Transformer: tokens and prompts

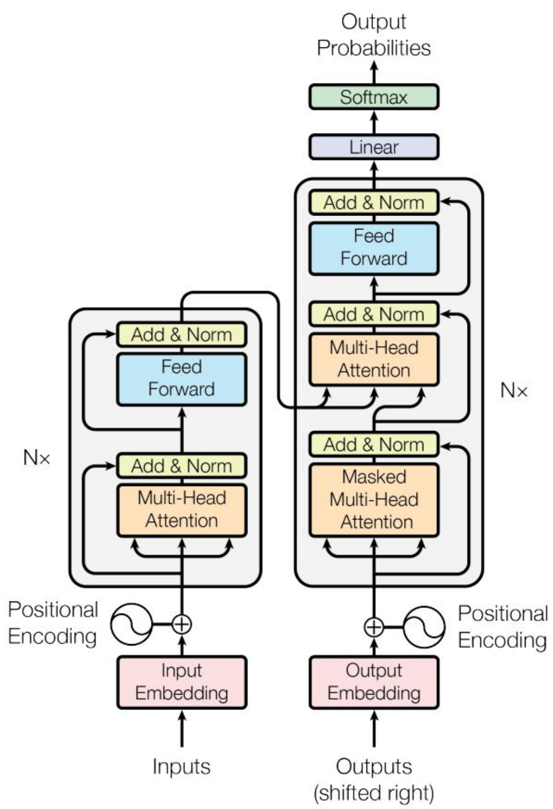


- Each $x_i = x_i(0) \in \mathbb{R}^d$: **token**
- Sequence $\{x_1(0), \ldots, x_n(0)\} \subset \mathbb{R}^d$: **prompt**
- Image data?



We dispose of N such input data-points $\{x_i(0)\}_{i \in [n]} \subset \mathbb{R}^d$

The Transformer: architecture



[Sander, Ablin, Blondel, Peyré '22]: given initial prompt $x_1(0), \ldots, x_n(0)$:

$$\dot{x}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{\langle Q x_i(t), K x_j(t) \rangle}}{\sum_{k=1}^{n} e^{\langle Q x_i(t), K x_k(t) \rangle}} \right) V x_j(t)$$

for $i \in [n]$.

Matrices (Q, K, V) are **controls**, and can be time-dependent.

Our goal

We are **given constant controls** (Q, K, V) (**trained Transformer**).

Question(s):

- What does the motion of the tokens $x_i(t)$ look like?
- Hidden geometric structure discovered by Transformers?
- How does it depend on (Q, K, V)?

What can we expect?

Take $Q^\top K = V = I_d$:

1. Emmanuel's talk: convergence to consensus: $x_i(t) \to \bar{x}_i$ and $\bar{x}_i = \bar{x}_j$

What can we expect?

Take $Q^\top K = V = I_d$:

1. Emmanuel's talk: convergence to consensus: $x_i(t) \to \bar{x}_i$ and $\bar{x}_i = \bar{x}_j$
2. Consider law of tokens $\mu(t,\cdot) \in \mathcal{P}_c(\mathbb{R}^d)$ (or empirical measure $\mu(t,\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)}$):

$$\partial_t \mu(t,x) + \operatorname{div}\left(\nabla \log\left(\int_{\mathbb{R}^d} e^{\langle x, x'\rangle} d\mu(t,x')\right) \mu(t,x)\right) = 0.$$

What can we expect?

Take $Q^\top K = V = I_d$:

1. Emmanuel's talk: convergence to consensus: $x_i(t) \to \bar{x}_i$ and $\bar{x}_i = \bar{x}_j$
2. Consider law of tokens $\mu(t,\cdot) \in \mathcal{P}_c(\mathbb{R}^d)$ (or empirical measure $\mu(t,\cdot) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$):

$$\partial_t \mu(t,x) + \operatorname{div}\left(\nabla \log\left(\int_{\mathbb{R}^d} e^{\langle x,x'\rangle} d\mu(t,x')\right)\mu(t,x)\right) = 0.$$

"Similarities" to Patlak-Keller-Segel:

$$\partial_t \mu(t,x) - \left(\nabla\left(\int_{\mathbb{R}} \log|x-x'| d\mu(t,x')\right)\mu(t,x)\right) = 0.$$

[Carrillo, Di Francesco, Figalli, Laurent, Slepcev '11]: $\exists T^*(\mu(0)) > 0$

$$\mu(t,x) = \delta_{\int x d\mu(0,x)} \qquad \text{for} \quad t \geqslant T^*.$$

Scope

1. **Results** (starting with $d = 1$, then $d > 1$)
2. **Proofs** (. . . are very low-technology)
3. **Beyond**

Results

- Assume $QK > 0$ and $V > 0$; WLOG $QK = V = 1$.

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{n} e^{x_i(t)x_k(t)}} \qquad (i,j) \in [n]^2.$$

- Assume $QK > 0$ and $V > 0$; WLOG $QK = V = 1$.

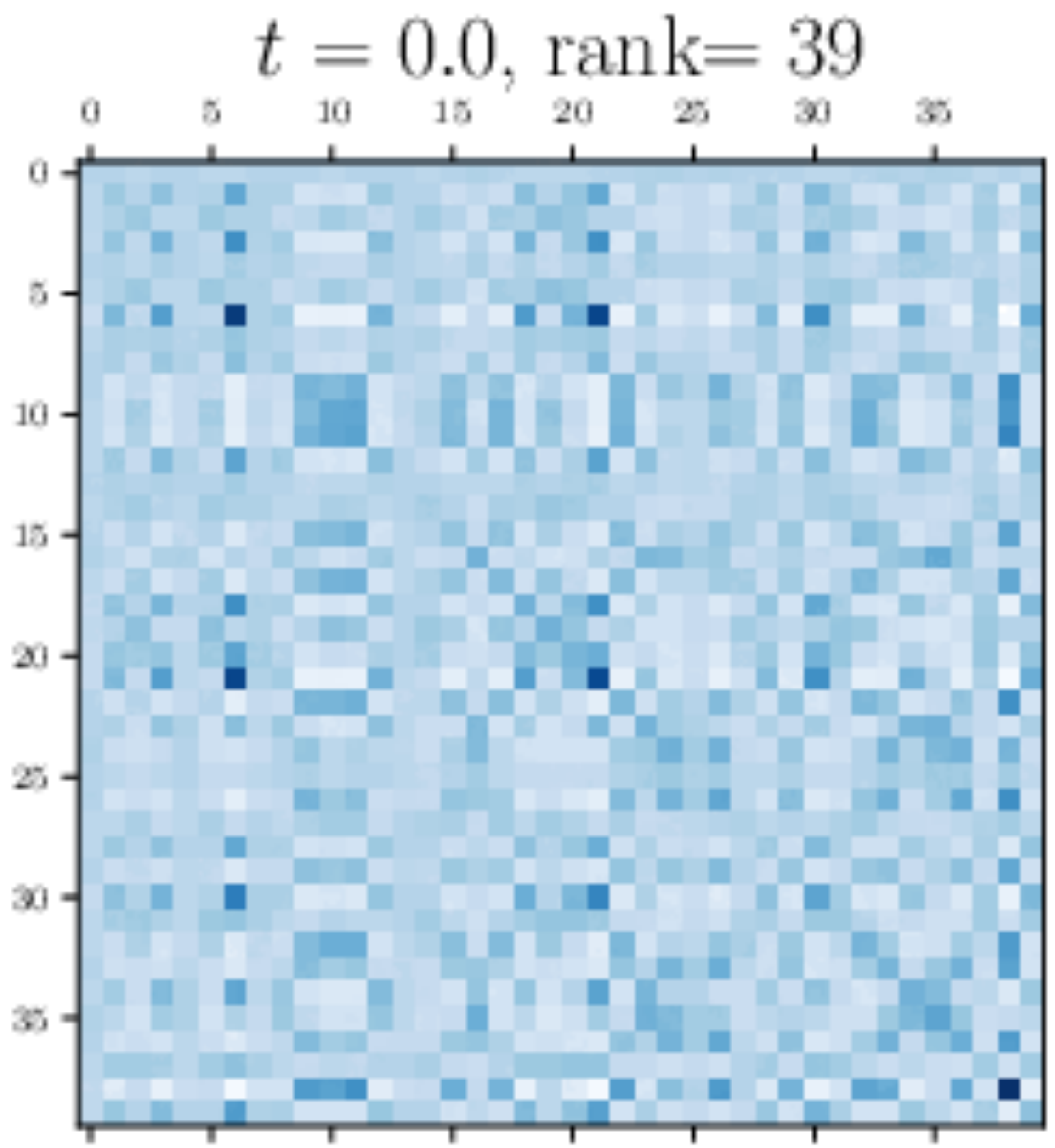
$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{n} e^{x_i(t)x_k(t)}} \qquad (i,j) \in [n]^2.$$

Theorem 1

Given prompt with $x_i(0) \neq x_j(0)$ for $i \neq j$. There exists $P^* \in \mathbb{R}^{n\times n}$:

$$P^* = \text{permutation}_1 \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ * & * & \dots & * \\ 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{bmatrix} \text{permutation}_2$$

s.t. $\lim_{t\to+\infty} P(t) = P^*$. Non-$*$ rows converge doubly exponentially fast.


$t = 0.0$, rank= 39
0
5
10
15
20
25
30
35

The case $d > 1$

$$\dot{x}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{\langle Qx_i(t), Kx_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle Qx_i(t), Kx_k(t)\rangle}} \right) Vx_j(t).$$

Suppose V has a positive eigenvalue.

The case $d > 1$

$$\dot{x}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{\langle Qx_i(t), Kx_j(t) \rangle}}{\sum_{k=1}^{n} e^{\langle Qx_i(t), Kx_k(t) \rangle}} \right) V x_j(t).$$

Suppose V has a positive eigenvalue.

Then $x_i(t)$ diverges to $\pm\infty$ with t, by analogy with

$$y'(t) = Vy(t).$$

The case $d > 1$

$$\dot{x}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{\langle Qx_i(t), Kx_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle Qx_i(t), Kx_k(t)\rangle}} \right) Vx_j(t).$$

Suppose V has a positive eigenvalue.

Then $x_i(t)$ diverges to $\pm\infty$ with t, by analogy with

$$y'(t) = Vy(t).$$

Change of time-scale

$$z_i(t) = e^{-tV} x_i(t)$$

Then

$$\dot{z}_i(t) = \sum_{j=1}^{n} \left(\frac{e^{\langle Qe^{tV} z_i(t), Ke^{tV} z_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle Qe^{tV} z_i(t), Ke^{tV} z_k(t)\rangle}} \right) V(z_j(t) - z_i(t))$$

$V = I_d$: Convex polytope

Theorem 2

Suppose $Q^\top K > 0$. There exists convex polytope $\mathscr{K} \subset \mathbb{R}^d$ of $m \geqslant 1$ vertices $v_1, \ldots, v_m$ such that for any $i \in [n]$,

$$\lim_{t \to \infty} z_i(t) = \overline{z}_i$$

for some $\overline{z}_i \in \partial\mathscr{K} \cup \{0\}$.

$V = I_d$: Convex polytope

Theorem 2

Suppose $Q^\top K > 0$. There exists convex polytope $\mathscr{K} \subset \mathbb{R}^d$ of $m \geqslant 1$ vertices $v_1, \ldots, v_m$ such that for any $i \in [n]$,

$$\lim_{t \to \infty} z_i(t) = \overline{z}_i$$

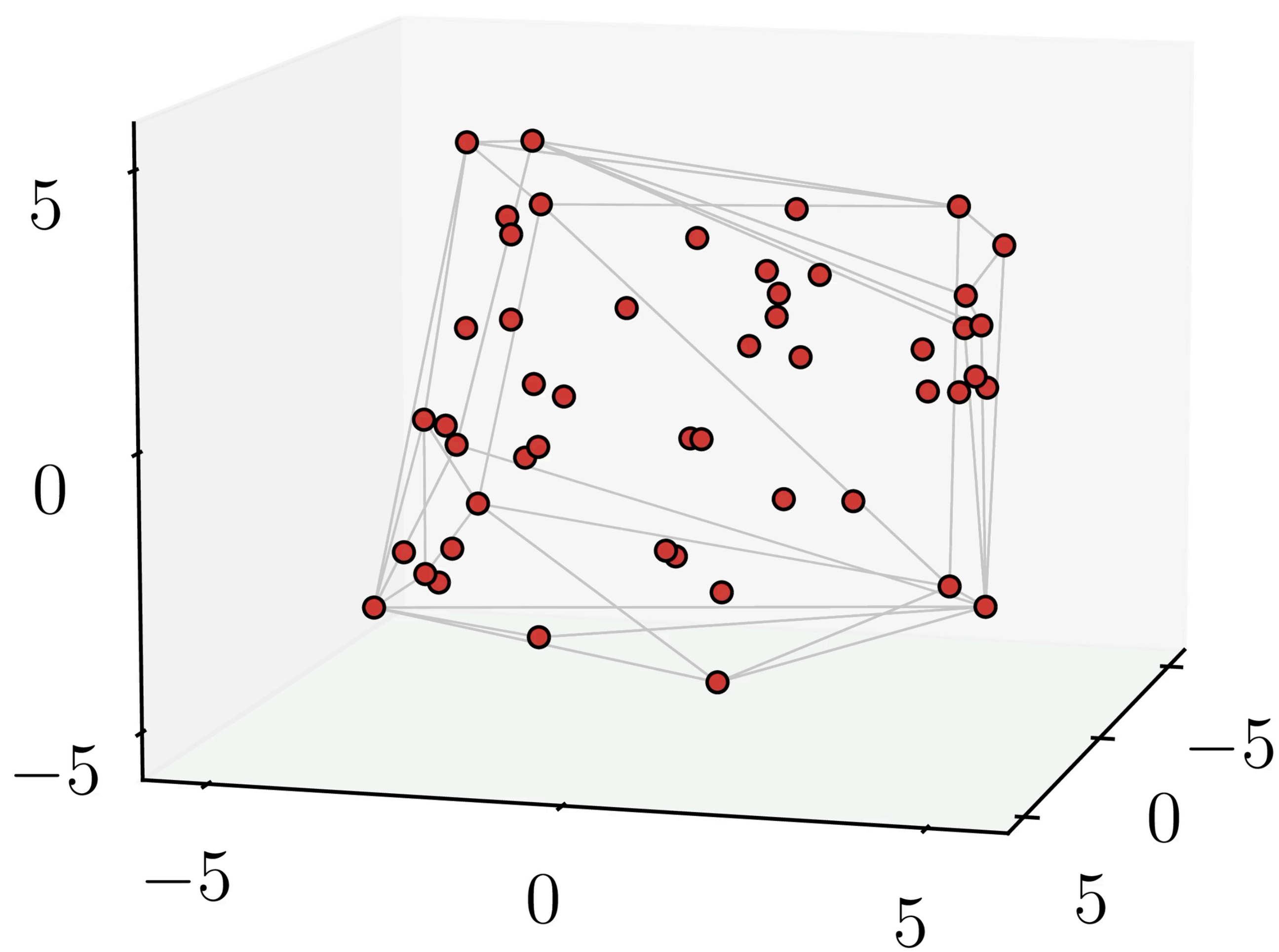
for some $\overline{z}_i \in \partial\mathscr{K} \cup \{0\}$. Actually $\overline{z}_i \in \mathcal{S}$ where

$$\{v_j\}_{j\in[m]} \subseteq \boxed{\mathcal{S} := \left\{ x \in \mathcal{K} \colon \|Ax\|^2 = \max_{j\in[m]} \langle Ax, Av_j \rangle \right\}} \subset \partial\mathscr{K} \cup \{0\}$$

and $A = (Q^\top K)^{\frac{1}{2}}$. $\mathcal{S}$ is discrete.

Corollary. $\lim_{t\to\infty} P(t) = P^*$, with rank $P^* \leqslant m$.

$t = 0.0$



Parallel hyperplanes

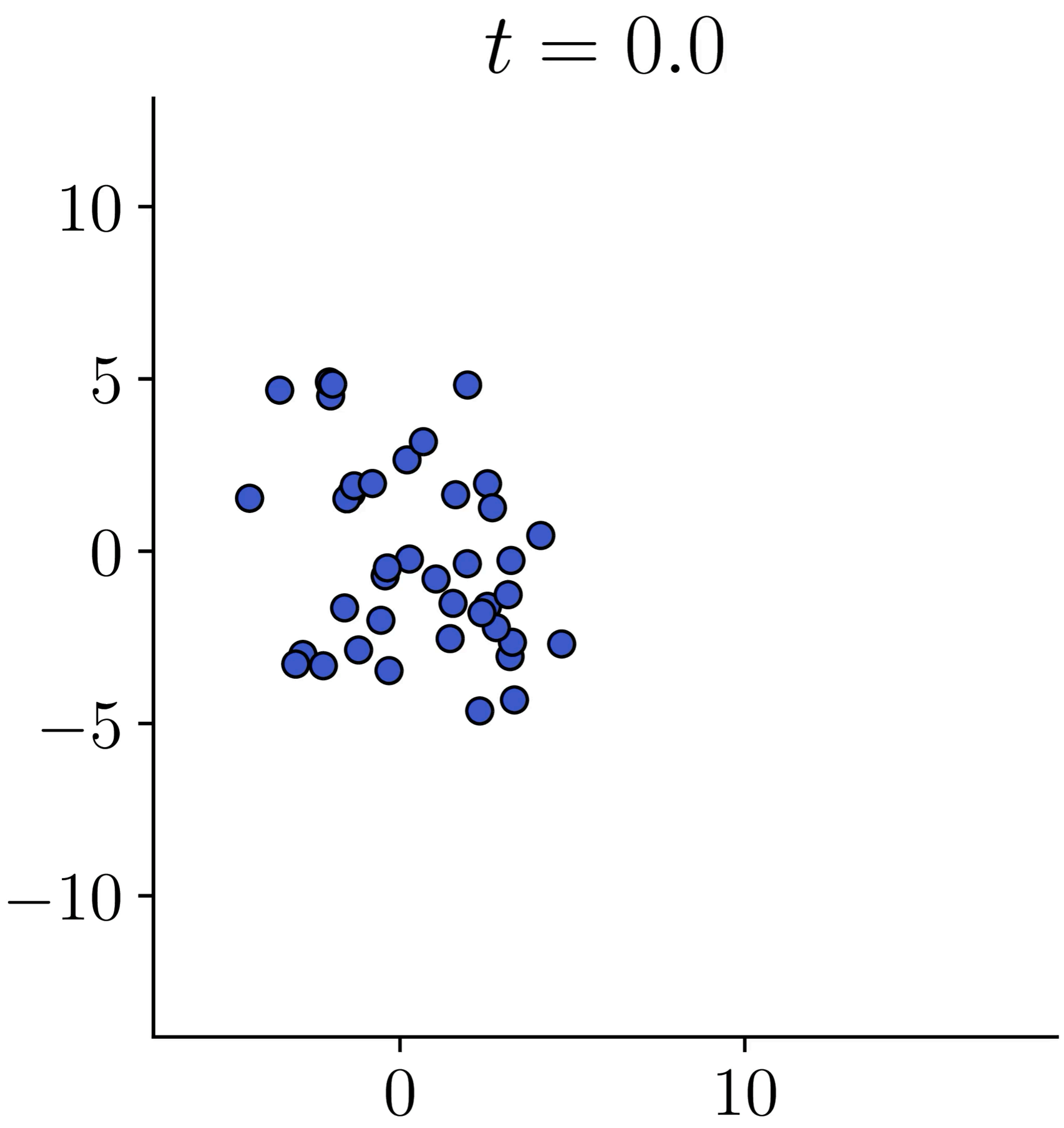
$V \in \mathbb{R}^{d\times d}$ such that

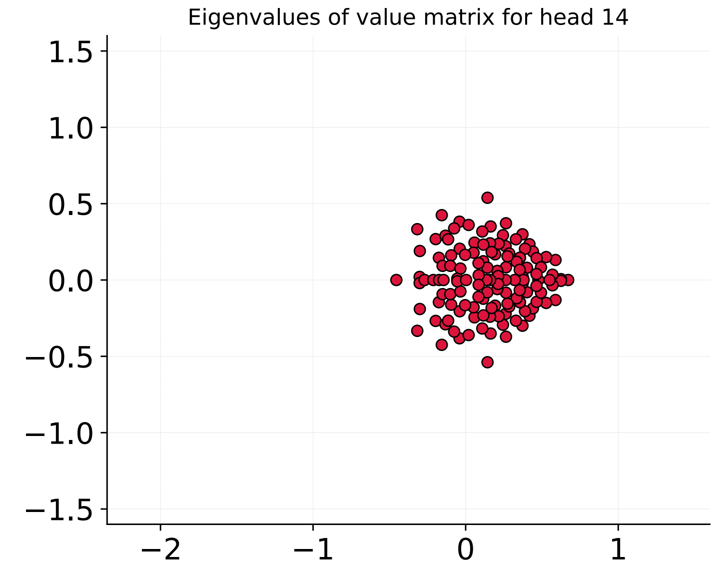
$$\lambda_1 := \max_{j\in[d]} |\lambda_j|$$

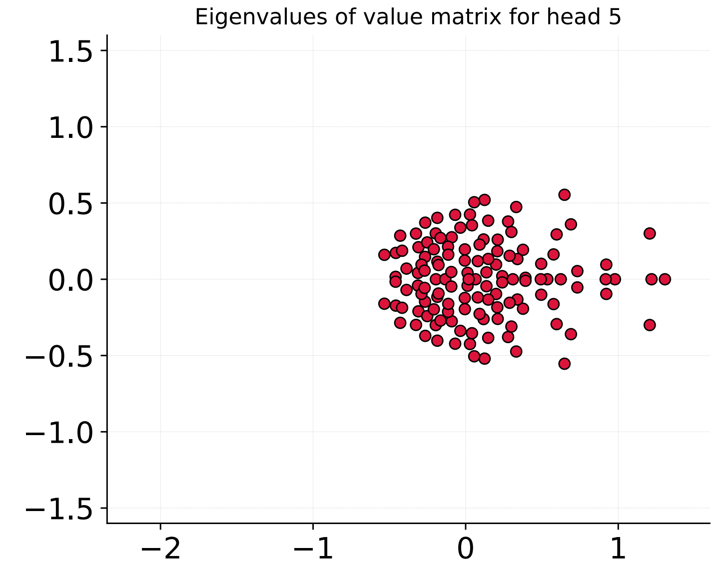
satisfies $\lambda_1 > 0$ and is simple (Perron-Frobenius).

Theorem 3

Suppose $Q^\top K \succ 0$. There exist (at most) three parallel hyperplanes such that for any $i \in [n]$, $z_i(t)$ converges to one of these hyperplanes as $t \to \infty$.


$t = 0.0$
10
5
0
−5
−10
0
10


Eigenvalues of value matrix for head 14
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1


Eigenvalues of value matrix for head 5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1

Polytopes × hyperplanes

Linear subspaces $F, G \subset \mathbb{R}^d$, both invariant under V, with

$$F \oplus G = \mathbb{R}^d$$

and

$$V_{|F} = \lambda I_d$$

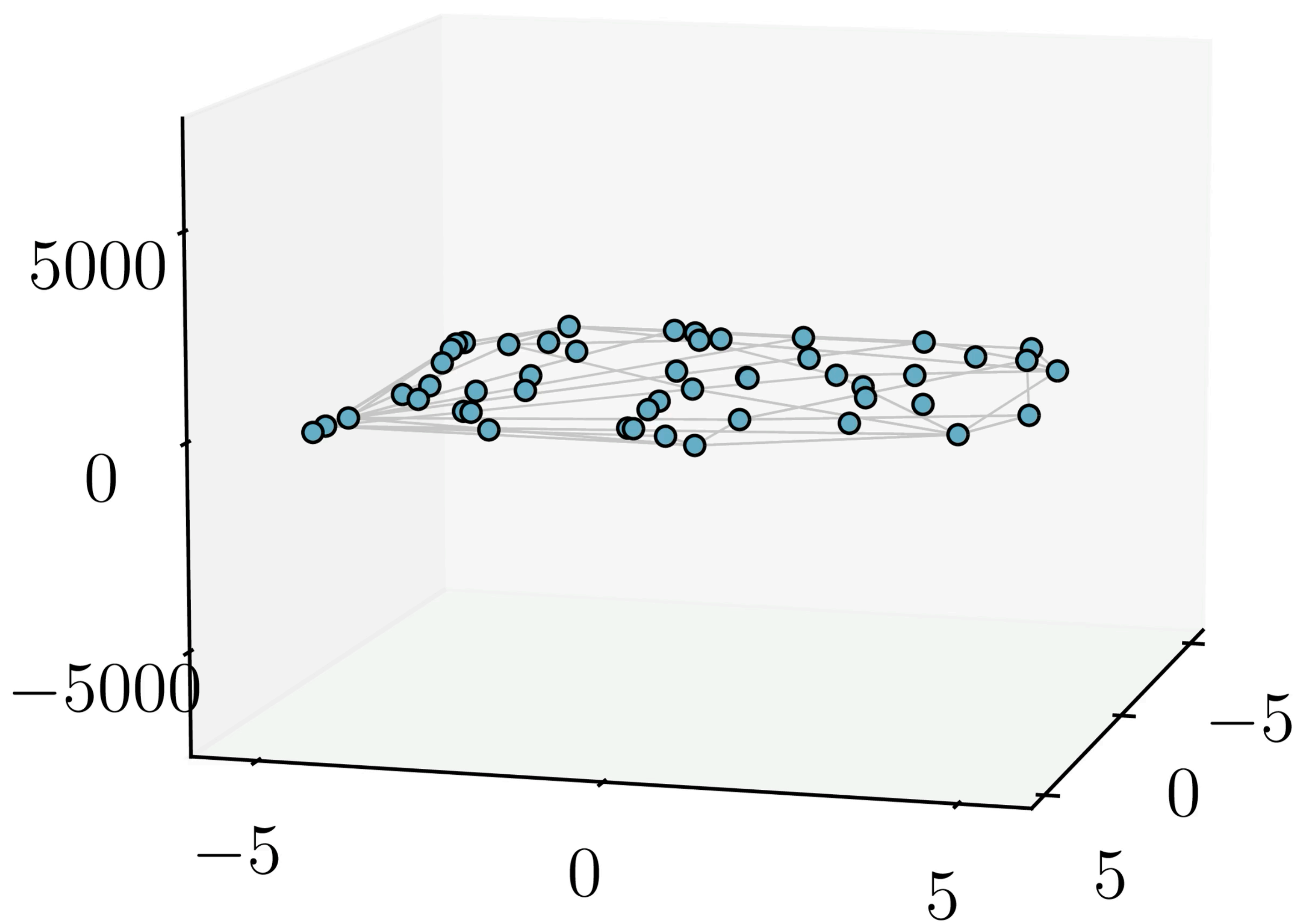
for some $\lambda > 0$, and

$$\max_j |\lambda_j(V_{|G})| < \lambda.$$

Theorem 4

Suppose $Q^\top K \succ 0$. There exists a bounded convex polytope $\mathscr{K} \subset F$ such that $z_i(t)$ converge to $\partial\mathscr{K} \times G$ as $t \to \infty$.

$t = 0.0$



Proofs

Proof of Theorem 1

Setup: $d = 1$, $QK = V = 1$.

Lemma (any $d \geqslant 1$)

$t \mapsto \|x_i(t) - x_j(t)\|$ is increasing for any $i, j \in [n]$.

Proof of Theorem 1

Setup: $d = 1$, $QK = V = 1$.

Lemma (any $d \geqslant 1$)

$t \mapsto \|x_i(t) - x_j(t)\|$ is increasing for any $i, j \in [n]$.

Proof. Have

$$\dot{x}_i(t) = \nabla f(x_i(t))$$

where $f(z) = \log \sum_{j=1}^{n} e^{\langle z, x_j \rangle}$ is convex. By convexity

$$\frac{1}{2}\frac{d}{dt}\|x_i - x_j\|^2 = \langle \nabla f(x_i) - \nabla f(x_j), x_i - x_j \rangle \geqslant 0. \qquad \square$$

Proof of Theorem 1

Setup: $d = 1$, $QK = V = 1$.

Lemma (any $d \geqslant 1$)

$t \mapsto \|x_i(t) - x_j(t)\|$ is increasing for any $i, j \in [n]$.

Proof. Have

$$\dot{x}_i(t) = \nabla f(x_i(t))$$

where $f(z) = \log \sum_{j=1}^{n} e^{\langle z, x_j \rangle}$ is convex. By convexity

$$\frac{1}{2}\frac{d}{dt}\|x_i - x_j\|^2 = \langle \nabla f(x_i) - \nabla f(x_j), x_i - x_j \rangle \geqslant 0. \quad \square$$

So particles are growing apart. What if $\lim_{t\to\infty} x_i(t) = \pm\infty$?

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t \to \infty} x_i(t) = \infty$.

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t\to\infty} x_i(t) = \infty$. For any $j \neq n$,

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^n e^{x_i(t)x_k(t)}} = \frac{1}{\sum_{k=1}^n e^{x_i(t)(x_k(t)-x_j(t))}} \leqslant e^{-x_i(t)(x_n(t)-x_j(t))} \leqslant e^{-cx_i(t)}.$$

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t \to \infty} x_i(t) = \infty$. For any $j \neq n$,

$$P_{ij}(t) = \frac{e^{x_i(t) x_j(t)}}{\sum_{k=1}^n e^{x_i(t) x_k(t)}} = \frac{1}{\sum_{k=1}^n e^{x_i(t)(x_k(t) - x_j(t))}}$$
$$\leqslant e^{-x_i(t)(x_n(t) - x_j(t))} \leqslant e^{-c x_i(t)}.$$

All but the last component of i-th row of $P(t)$ go to 0.

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t\to\infty} x_i(t) = \infty$. For any $j \neq n$,

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^n e^{x_i(t)x_k(t)}} = \frac{1}{\sum_{k=1}^n e^{x_i(t)(x_k(t)-x_j(t))}} \leqslant e^{-x_i(t)(x_n(t)-x_j(t))} \leqslant e^{-cx_i(t)}.$$

All but the last component of i-th row of $P(t)$ go to 0. Since $P(t)$ stochastic,

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t\to\infty} x_i(t) = \infty$. For any $j \neq n$,

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^n e^{x_i(t)x_k(t)}} = \frac{1}{\sum_{k=1}^n e^{x_i(t)(x_k(t)-x_j(t))}} \leqslant e^{-x_i(t)(x_n(t)-x_j(t))} \leqslant e^{-cx_i(t)}.$$

All but the last component of i-th row of $P(t)$ go to 0. Since $P(t)$ stochastic,

$$P_{in}(t) = 1 - \sum_{j=1}^{n-1} P_{ij}(t) \to 1.$$

So if $\lim_{t\to\infty} x_i(t) = +\infty$, then $P_i(t) \to e_n$.

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i \in [n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t\to\infty} x_i(t) = \infty$. For any $j \neq n$,

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^n e^{x_i(t)x_k(t)}} = \frac{1}{\sum_{k=1}^n e^{x_i(t)(x_k(t)-x_j(t))}} \leqslant e^{-x_i(t)(x_n(t)-x_j(t))} \leqslant e^{-cx_i(t)}.$$

All but the last component of i-th row of $P(t)$ go to 0. Since $P(t)$ stochastic,

$$P_{in}(t) = 1 - \sum_{j=1}^{n-1} P_{ij}(t) \to 1.$$

So if $\lim_{t\to\infty} x_i(t) = +\infty$, then $P_i(t) \to e_n$.

Similarly if $\lim_{t\to\infty} x_i(t) = -\infty$ then $P_i(t) \to e_1$.

WLOG particles are ordered: $x_1(t) \leqslant \ldots \leqslant x_n(t)$,

$$c := \min_{i\in[n-1]} (x_{i+1}(0) - x_i(0)) > 0.$$

Suppose $\lim_{t\to\infty} x_i(t) = \infty$. For any $j \neq n$,

$$P_{ij}(t) = \frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{n} e^{x_i(t)x_k(t)}} = \frac{1}{\sum_{k=1}^{n} e^{x_i(t)(x_k(t)-x_j(t))}} \leqslant e^{-x_i(t)(x_n(t)-x_j(t))} \leqslant e^{-cx_i(t)}.$$

All but the last component of i-th row of $P(t)$ go to 0. Since $P(t)$ stochastic,

$$P_{in}(t) = 1 - \sum_{j=1}^{n-1} P_{ij}(t) \to 1.$$

So if $\lim_{t\to\infty} x_i(t) = +\infty$, then $P_i(t) \to e_n$.

Similarly if $\lim_{t\to\infty} x_i(t) = -\infty$ then $P_i(t) \to e_1$.

Not easy:

- all but one particle tend to $\pm\infty$
- if $x_1(t)$ or $x_n(t)$ bounded, still get $P_1(t) \to e_1$ or $P_n(t) \to e_n$
- if internal particle is bounded, then we get the $*$-row.

Proof of Theorem 2

Setup: $V = I_d$, $Q^\top K > 0$; working with $z_i(t)$.

Lemma

$t \mapsto \mathrm{conv}\{z_i(t)\}_{i\in[n]}$ is decreasing:

$$\mathrm{conv}\{z_i(t_2)\}_{i\in[n]} \subseteq \mathrm{conv}\{z_i(t_1)\}_{i\in[n]}$$

if $t_1 \leqslant t_2$.

Inspired by [Jabin, Motsch '14] (opinion dynamics).

Proof of Theorem 2

Setup: $V = I_d$, $Q^\top K \succ 0$; working with $z_i(t)$.

Lemma

$t \mapsto \mathrm{conv}\{z_i(t)\}_{i\in[n]}$ is decreasing:

$$\mathrm{conv}\{z_i(t_2)\}_{i\in[n]} \subseteq \mathrm{conv}\{z_i(t_1)\}_{i\in[n]}$$

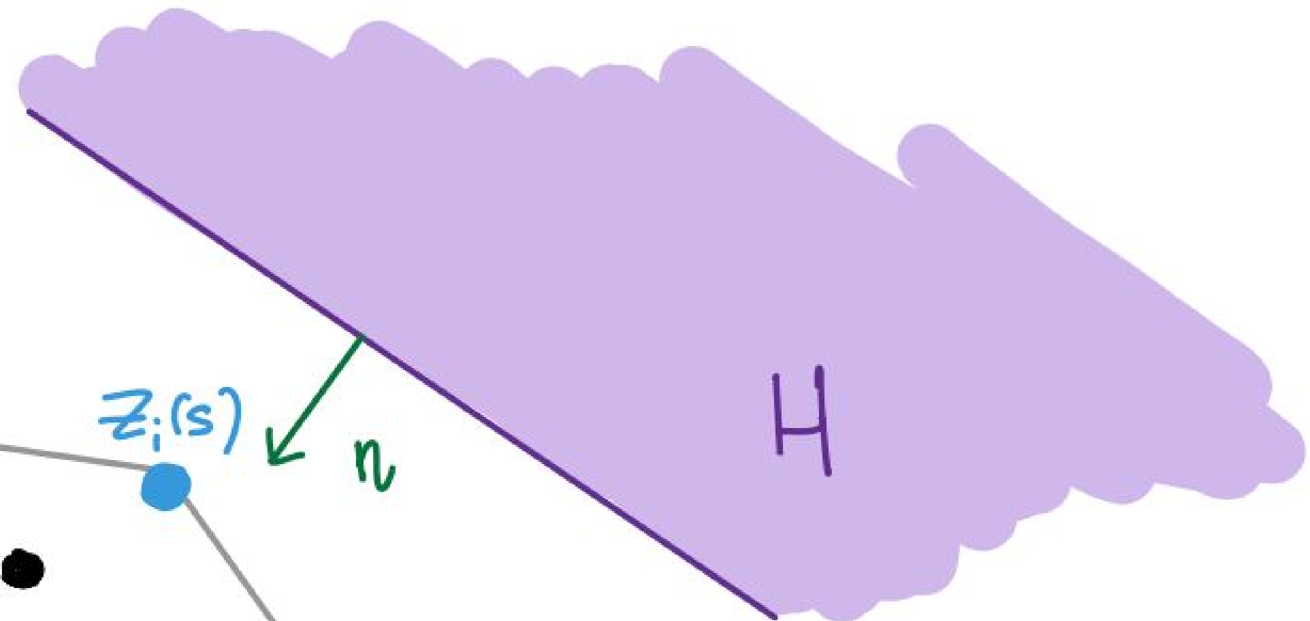
if $t_1 \leqslant t_2$.

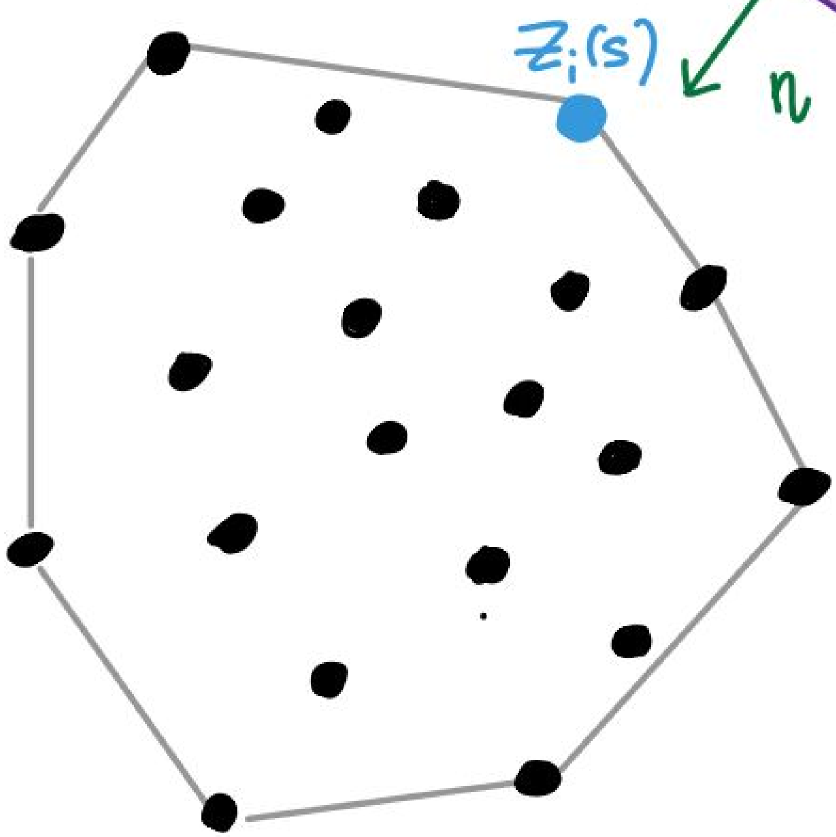
Inspired by [Jabin, Motsch '14] (opinion dynamics).

Proof. Fix $t > 0$. Let $H \subset \mathbb{R}^d$ be closed half-space not containing any $z_i(t)$. Then

$$\alpha : s \mapsto \min_{i\in[n]} \mathrm{dist}(z_i(s), H)$$

is increasing.



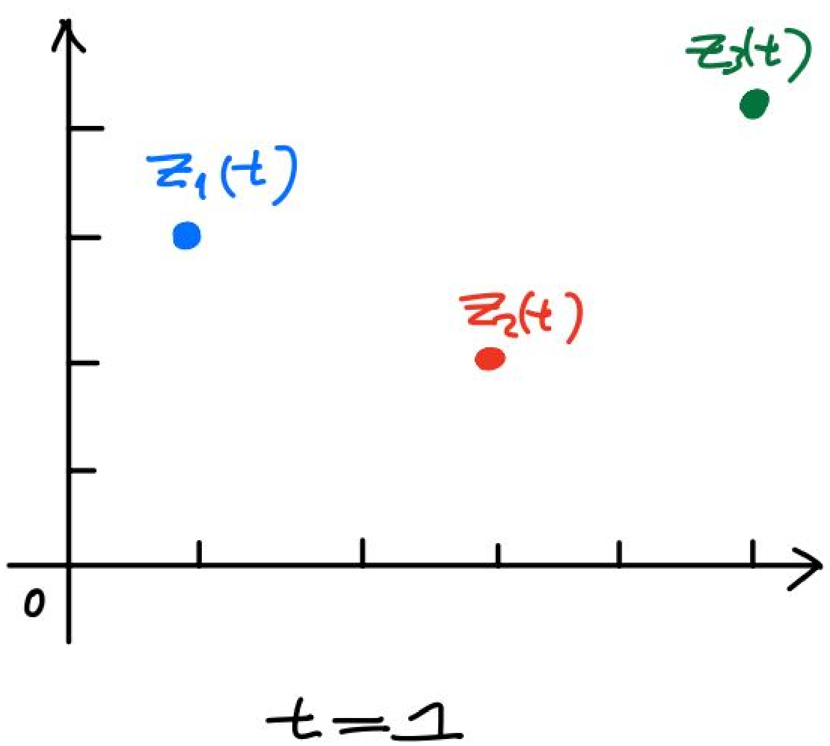


$$\frac{d}{dt}\operatorname{dist}(z_i(s), H) = \langle \dot{z}_i(s), n \rangle = \sum_{j=1}^{n} \underbrace{P_{ij}(s)}_{\geq 0.} \underbrace{\langle z_j(s) - z_i(s), n \rangle}_{\geq 0}$$

Two timescales

1. During $t \sim O(1)$, we follow the Lemma.
2. Once $t = O(1)$: $e^{2t} = \beta$ is gigantic, and

$$\sum_{j=1}^{n} \left(\frac{e^{\beta\langle Az_i(t), Az_j(t)\rangle}}{\sum_{k=1}^{n} e^{\beta\langle Az_i(t), Az_k(t)\rangle}} \right) (z_j(t) - z_i(t)) \approx \sum_{j \in \mathrm{argmax}_{k\in[n]}\langle Az_i(t), Az_k(t)\rangle} (z_j(t) - z_i(t)).$$

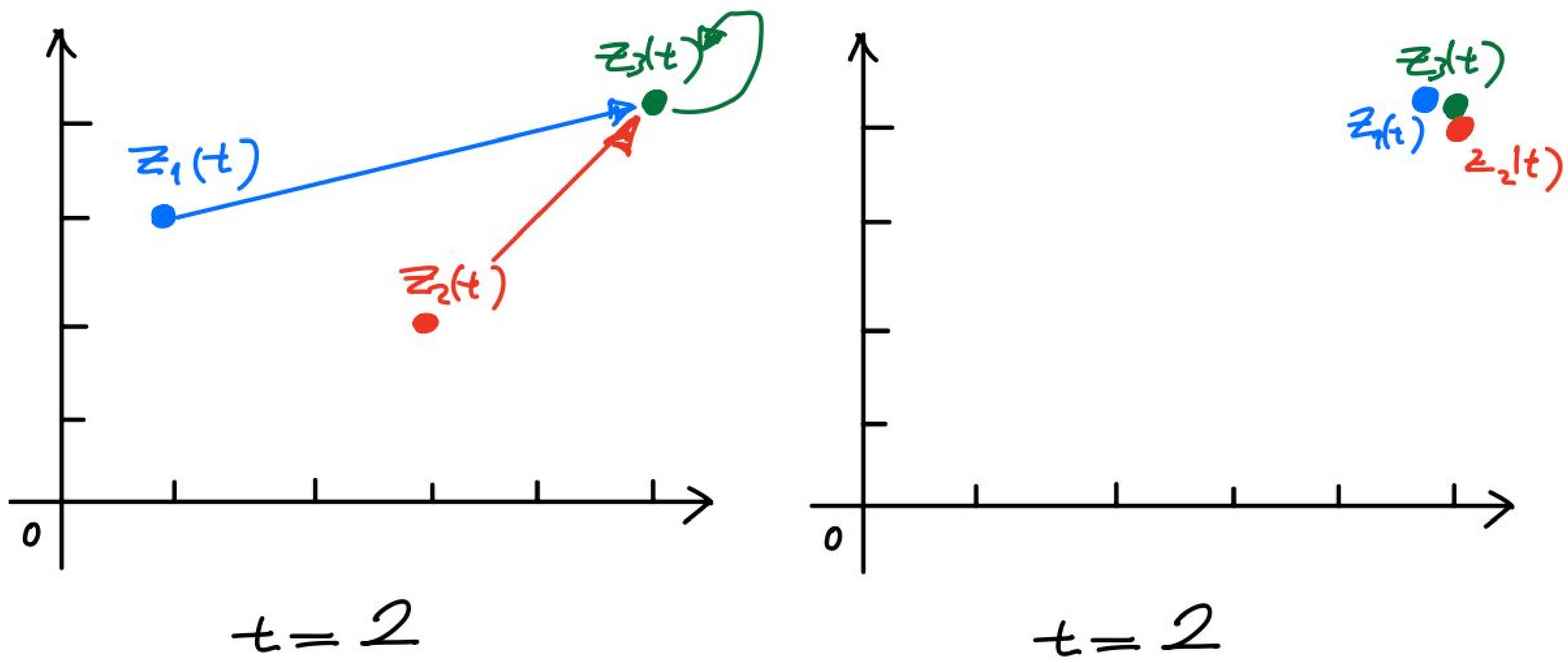


Two timescales

1. During $t \sim O(1)$, we follow the Lemma.
2. Once $t = O(1)$: $e^{2t} = \beta$ is gigantic, and

$$\sum_{j=1}^{n} \left(\frac{e^{\beta\langle Az_i(t), Az_j(t)\rangle}}{\sum_{k=1}^{n} e^{\beta\langle Az_i(t), Az_k(t)\rangle}} \right) (z_j(t) - z_i(t))$$

$$\approx \sum_{j \in \mathrm{argmax}_{k\in[n]}\langle Az_i(t), Az_k(t)\rangle} (z_j(t) - z_i(t)).$$



Proof of Theorem 3

Setup: $\lambda_1 > 0$ simple, $Q^\top K \succ 0$, working with $z_i(t)$. V diagonalizable.

Lemma

Let $k \in [d]$ s.t. $\lambda_k \geqslant 0$. Then

$$a_k : t \mapsto \min_{j \in [n]} \varphi_k^*(z_j(t))$$

is increasing, and

$$b_k : t \mapsto \max_{j \in [n]} \varphi_k^*(z_j(t))$$

is decreasing. (Here $\varphi_1^*, \ldots, \varphi_d^*$ dual basis.)

Proof of Theorem 3

Setup: $\lambda_1 > 0$ simple, $Q^\top K \succ 0$, working with $z_i(t)$. V diagonalizable.

Lemma

Let $k \in [d]$ s.t. $\lambda_k \geqslant 0$. Then

$$a_k : t \mapsto \min_{j \in [n]} \varphi_k^*(z_j(t))$$

is increasing, and

$$b_k : t \mapsto \max_{j \in [n]} \varphi_k^*(z_j(t))$$

is decreasing. (Here $\varphi_1^*, \ldots, \varphi_d^*$ dual basis.)

Proof. Let $i \in [n]$ s.t. $a_k(t) = \varphi_k^*(z_i(t))$. Then

$$\frac{d}{dt}\varphi_k^*(z_i(t)) = \sum_{j=1}^n P_{ij}\varphi_k^*(V(z_j(t) - z_i(t))) = \lambda_k \sum_{j=1}^n P_{ij}(\varphi_k^*(z_j(t)) - \varphi_k^*(z_i(t))).$$

This is $\geqslant 0$ by $\lambda_k \geqslant 0$ and choice of i. □

We show that

$$\lim_{t\to\infty} \varphi_1^*(z_i(t)) = c$$

where $c \in \{0, a, b\}$, and

$$a = \lim_{t\to\infty} \min_{j\in[n]} \varphi_1^*(z_j(t)), \qquad b = \lim_{t\to\infty} \max_{j\in[n]} \varphi_1^*(z_j(t)).$$

So, three parallel hyperplanes are $H_c = \{z \colon \varphi_1^*(z) = c\}$.

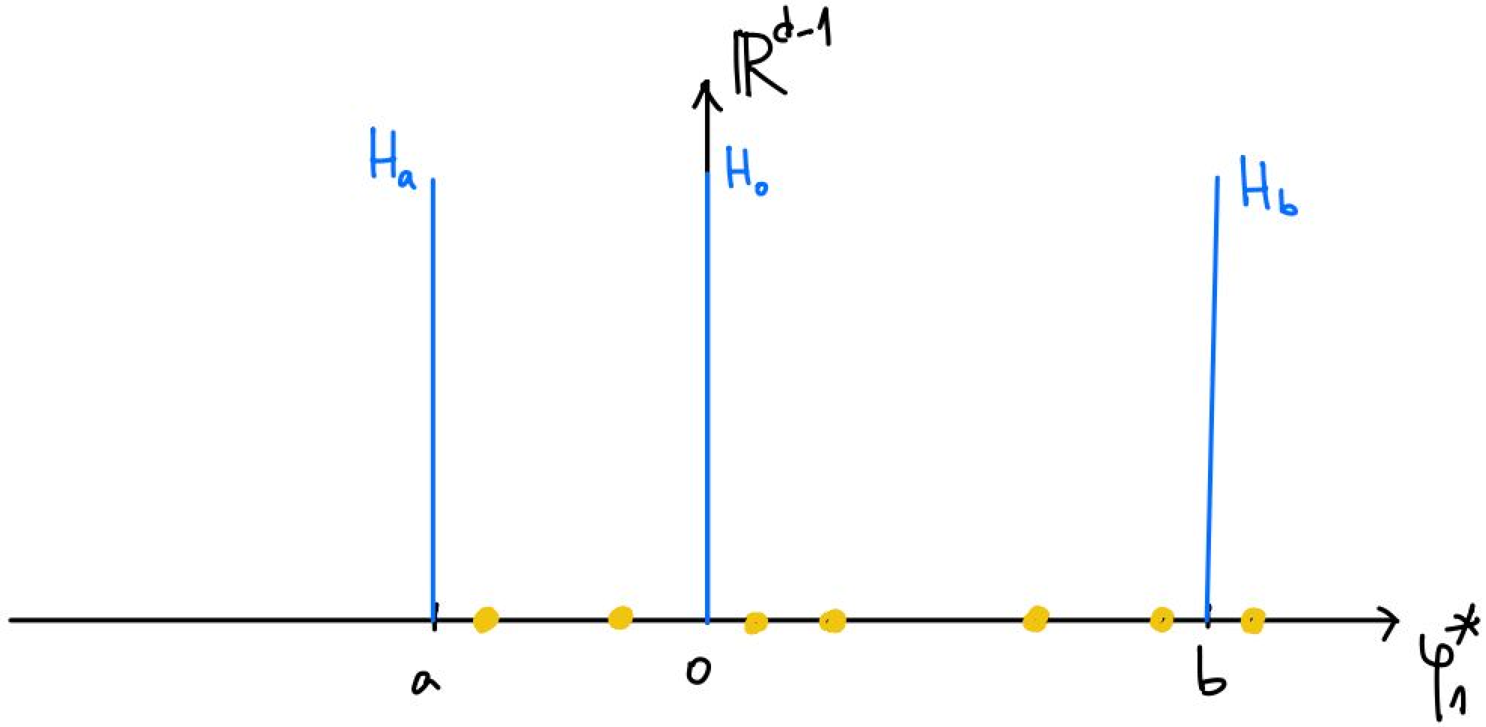
We show that

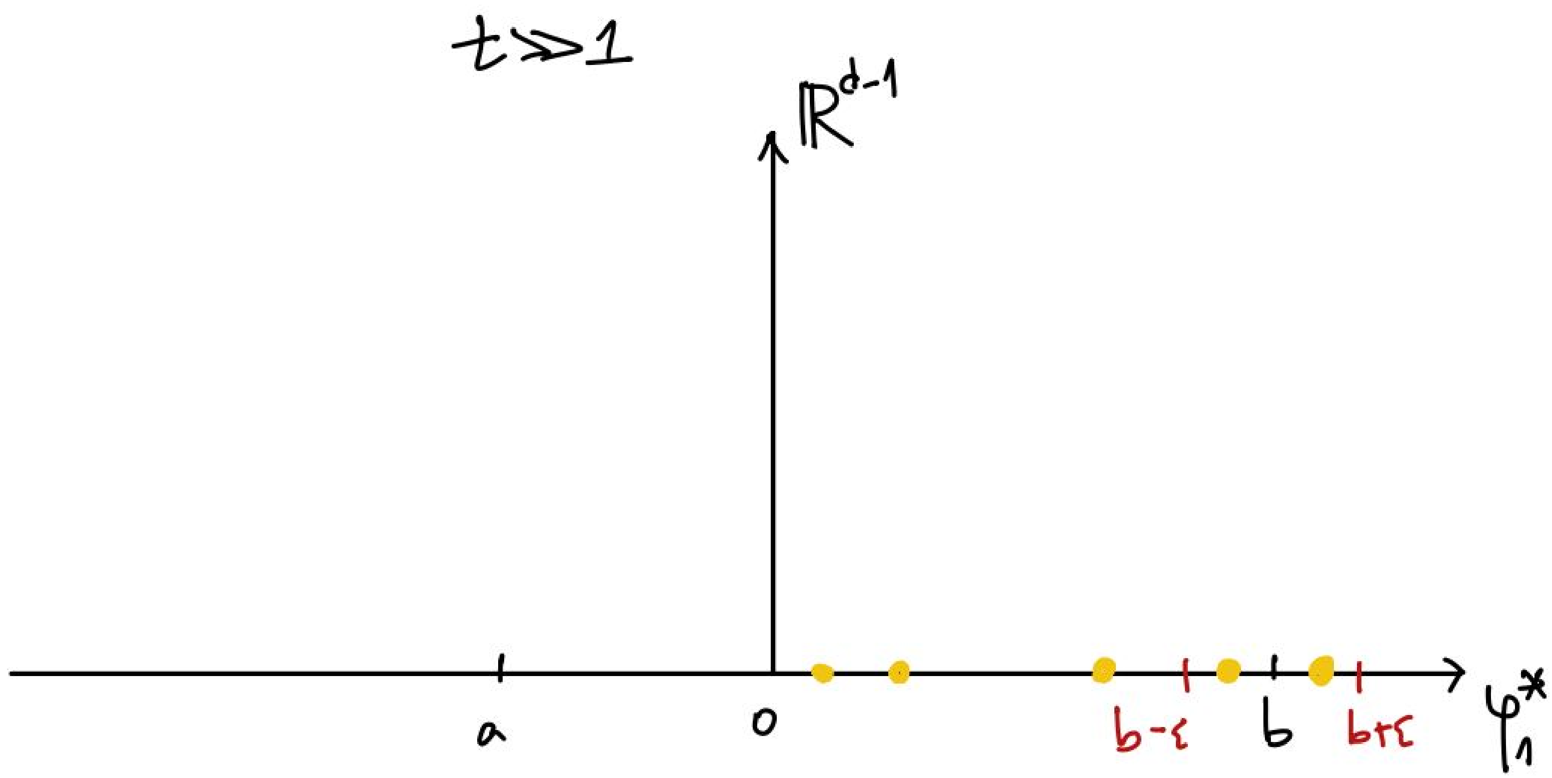
$$\lim_{t\to\infty} \varphi_1^*(z_i(t)) = c$$

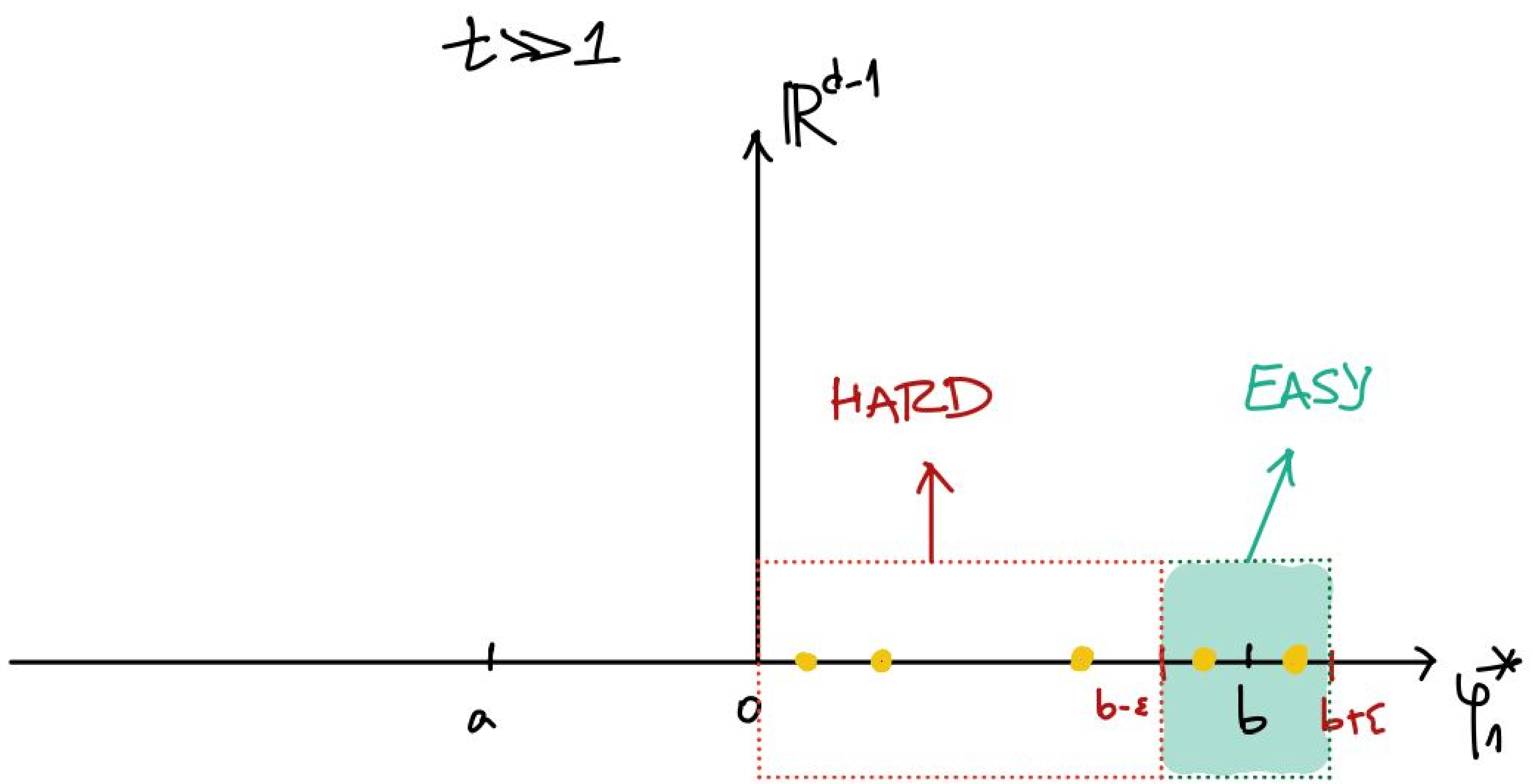
where $c \in \{0, a, b\}$, and

$$a = \lim_{t\to\infty} \min_{j\in[n]} \varphi_1^*(z_j(t)), \qquad b = \lim_{t\to\infty} \max_{j\in[n]} \varphi_1^*(z_j(t)).$$

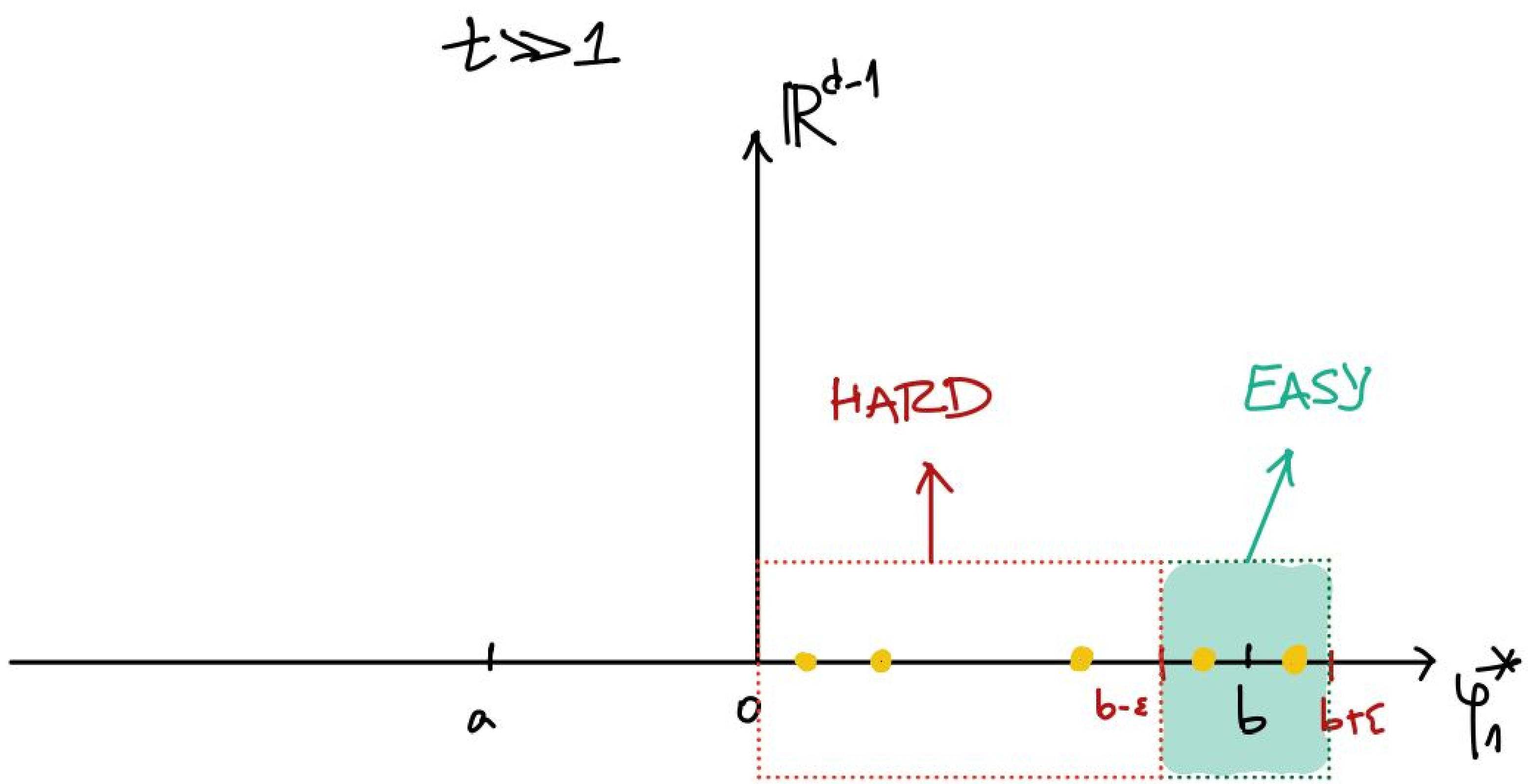
So, three parallel hyperplanes are $H_c = \{z : \varphi_1^*(z) = c\}$.




t≫1
ℝ^{d-1}
a
0
b-ε
b
b+ε
φ*_1

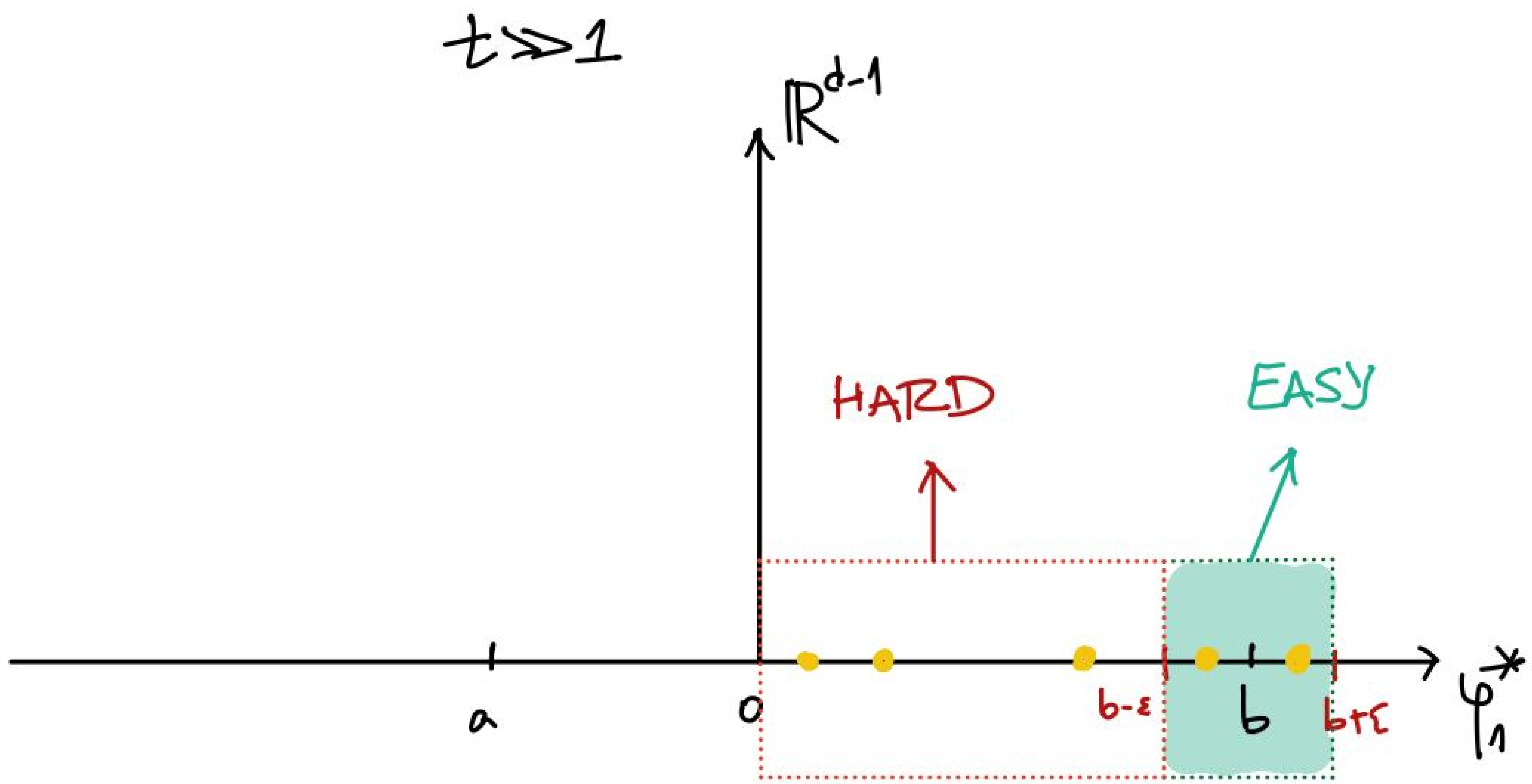


Hard regime: $\forall \varepsilon \ll 1$, suppose $\varphi_1^*(z_i(t)) \in [\varepsilon, b-\varepsilon]$ for t large enough.



Hard regime: $\forall \varepsilon \ll 1$, suppose $\varphi_1^*(z_i(t)) \in [\varepsilon, b-\varepsilon]$ for t large enough.

Goal. Show that $\varphi_1^*(z_i(t)) \geqslant b - \varepsilon$ for t even larger.



Hard regime: $\forall \varepsilon \ll 1$, suppose $\varphi_1^*(z_i(t)) \in [\varepsilon, b-\varepsilon]$ for t large enough.

Goal. Show that $\varphi_1^*(z_i(t)) \geqslant b-\varepsilon$ for t even larger.

How? Positive lower bound for

$$\frac{1}{\lambda_1}\frac{d}{dt}\varphi_1^*(z_i(t)) = \sum_{j=1}^{n} \frac{e^{w_j(t)}}{\sum_{k=1}^{n} e^{w_k(t)}}(\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t))).$$

Split the sum:

$$\frac{1}{\lambda_1}\frac{d}{dt}\varphi_1^*(z_i(t)) \geqslant \frac{e^{w_{j_0(t)}}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_{j_0(t)}(t)) - \varphi_1^*(z_i(t))\right)$$

$$+ \sum_{\{j:\, \varphi_1^*(z_j(t)) \leqslant\, \varphi_1^*(z_i(t))\}} \frac{e^{w_j(t)}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t))\right).$$

Split the sum:

$$\frac{1}{\lambda_1}\frac{d}{dt}\varphi_1^*(z_i(t)) \geqslant \frac{e^{w_{j_0(t)}}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_{j_0(t)}(t)) - \varphi_1^*(z_i(t))\right)$$

$$+ \sum_{\{j:\, \varphi_1^*(z_j(t))\leqslant\, \varphi_1^*(z_i(t))\}} \frac{e^{w_j(t)}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t))\right).$$

Recall:

$$w_j(t) = \langle e^{tV} z_i(t), e^{tV} z_j(t)\rangle = \sum_{k\neq\ell} e^{(\lambda_k+\lambda_\ell)t}\varphi_k^*(e^{tV} z_i(t))\varphi_\ell^*(e^{tV} z_j(t)).$$

Split the sum:

$$\frac{1}{\lambda_1}\frac{d}{dt}\varphi_1^*(z_i(t)) \geqslant \frac{e^{w_{j_0(t)}}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_{j_0(t)}(t)) - \varphi_1^*(z_i(t))\right)$$

$$+ \sum_{\{j\,:\,\varphi_1^*(z_j(t))\leqslant\,\varphi_1^*(z_i(t))\}} \frac{e^{w_j(t)}}{\sum_{k=1}^n e^{w_k(t)}}\left(\varphi_1^*(z_j(t)) - \varphi_1^*(z_i(t))\right).$$

Recall:

$$w_j(t) = \langle e^{tV}z_i(t), e^{tV}z_j(t)\rangle = \sum_{k\neq\ell} e^{(\lambda_k+\lambda_\ell)t}\varphi_k^*(e^{tV}z_i(t))\varphi_\ell^*(e^{tV}z_j(t)).$$

Use

$$|\varphi_k^*(e^{tV}z_i(t))| \leqslant Ce^{|\lambda_k|t}$$

to get

$$\boxed{\left|w_j(t) - e^{2\lambda_1 t}\varphi_1^*(z_i(t))\varphi_1^*(z_j(t))\right| \leqslant Ce^{(\lambda_1+|\lambda_2|)t}}$$

for $j \in [n]$.

Roughly speaking, $w_{j_0(t)}$ will be gigantic in front of all other terms.

Take-away

- Tokens converge to cluster geometries (which are strongly determined by V)
- Possible consequences on the rank of self-attention matrix $P(t)$
- Of interest due to possible reduction of $O(n^2)$ complexity of self-attention at every layer t

Beyond

Beyond

Results are still incomplete for this model:

- Beyond Theorem 2 (polytope): convergence to vertices (and count them!) for positive measure set of initial conditions?

Beyond

Results are still incomplete for this model:

- Beyond Theorem 2 (polytope): convergence to vertices (and count them!) for positive measure set of initial conditions? Geometric description of polytope for generic data?

Beyond

Results are still incomplete for this model:

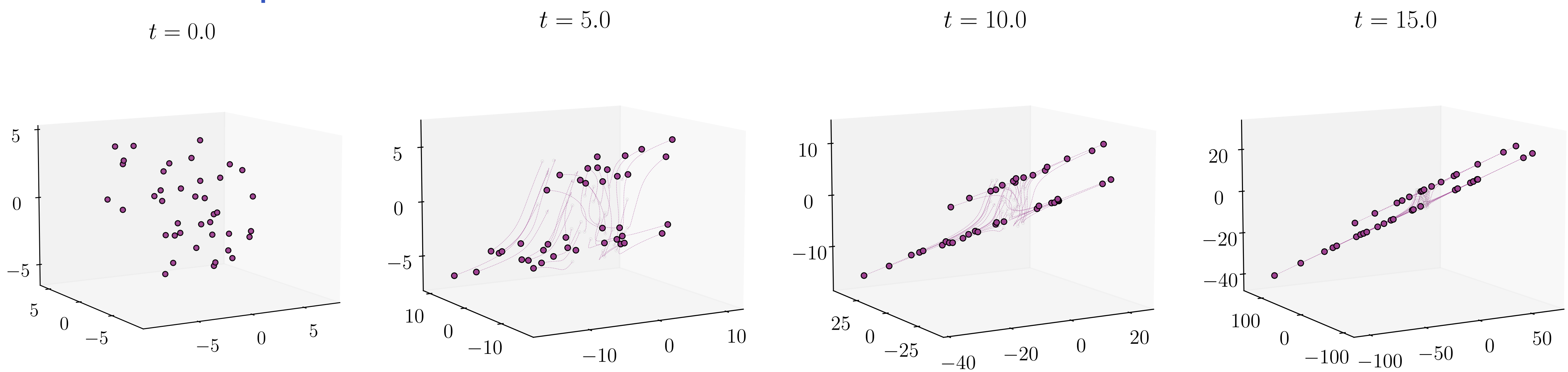
- Beyond Theorem 2 (polytope): convergence to vertices (and count them!) for positive measure set of initial conditions? Geometric description of polytope for generic data?
- Beyond Theorem 3 (hyperplanes): if $\mathbf{Re}(\lambda_j) \geqslant 0$ for k eigenvalues of V, then convergence to at most 3 parallel codimension-k subspaces?

Beyond

Results are still incomplete for this model:

- Beyond Theorem 2 (polytope): convergence to vertices (and count them!) for positive measure set of initial conditions? Geometric description of polytope for generic data?

- Beyond Theorem 3 (hyperplanes): if $\mathbf{Re}(\lambda_j) \geqslant 0$ for k eigenvalues of V, then convergence to at most 3 parallel codimension-k subspaces?



- What about rank $P(t)$ in general?