

Anisotropic interaction in self-attention layers

Workshop on the Mathematics of Transformers,
Oxford

Samira Kabri, Deutsches Elektronen-Synchrotron DESY, Hamburg

August 21, 2026

Joint work with **Martin Burger** (DESY, UHH), **Yury Korolev** (Uni Bath),
Tim Roith (TU Munich), **Lukas Weigand** (DESY)

Language processing: What is *it*?



"The cat did not see the moon because **it** did not open its eyes."



Language processing: What is *it*?



"The cat did not see the moon because **it** was hidden behind a cloud."

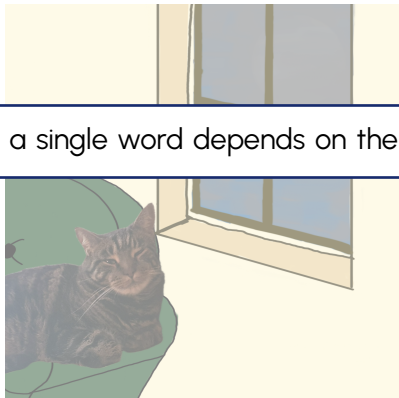


Language processing: What is *it*?



"The cat did not see the moon because *it* was hidden behind a cloud."

The meaning of a single word depends on the whole sentence.

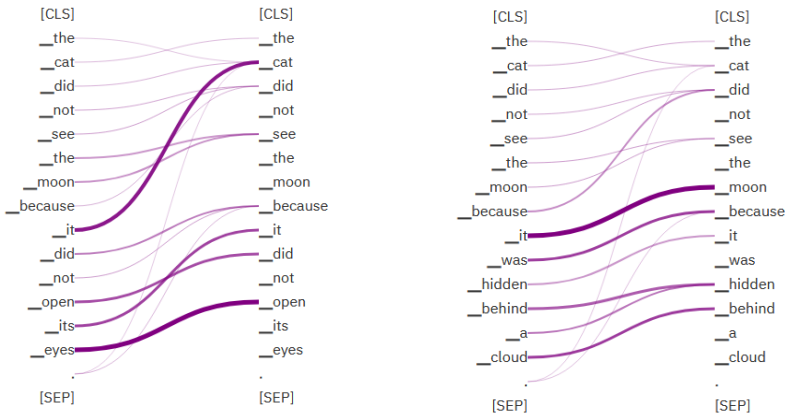


Language processing: What is *it*?



Deutsche
Forschungsgemeinschaft

Relations between words based on **attention matrices**:



Source: exBERT [HSG], model *albert-base-v1*, layer 11, head 2

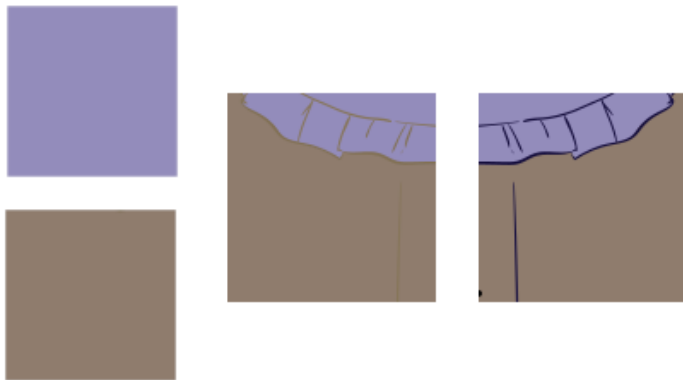
August 21, 2026

Self-attention dynamics

samira.kabri@desy.de

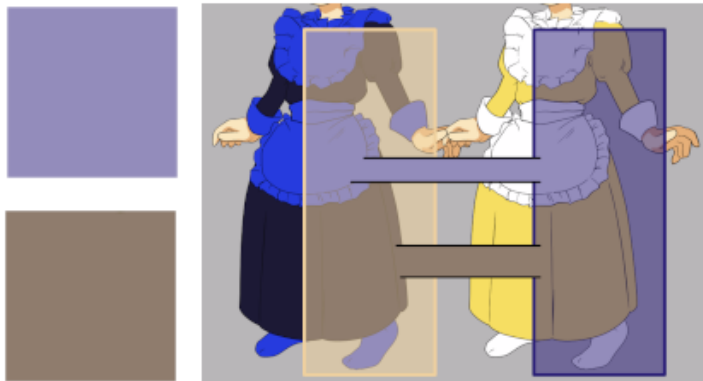
2/24

A famous imaging example



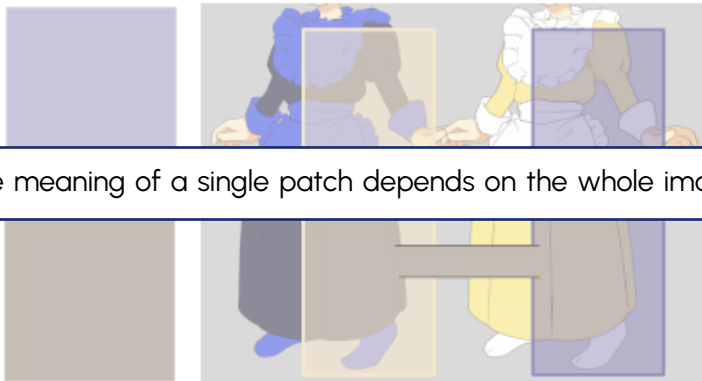
Original image: https://commons.wikimedia.org/wiki/File:Wikipet-an_wearing_The_Dress_reduced.svg

A famous imaging example



Original image: https://commons.wikimedia.org/wiki/File:Wikepe-tan_wearing_The_Dress_reduced.svg

A famous imaging example



Original image: https://commons.wikimedia.org/wiki/File:Wikepe-tan_wearing_The_Dress_reduced.svg



1. Transformer Architecture

2. Self-Attention Layers

3. Evolution of Measures

4. Conclusion and Outlook



Transformer Architecture

- Transformer **architecture**:

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention is all you need*. 2017

- Transformer **architecture**:
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention is all you need*. 2017
- **PDE view** on self-attention:
M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. *Sinkformers: Transformers with doubly stochastic attention*. 2022

- Transformer **architecture**:
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention is all you need*. 2017
- **PDE view** on self-attention:
M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. *Sinkformers: Transformers with doubly stochastic attention*. 2022
- Setting on the **sphere**:
B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. *A mathematical perspective on transformers*. 2025

- Transformer **architecture**:
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention is all you need*. 2017
- **PDE view** on self-attention:
M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. *Sinkformers: Transformers with doubly stochastic attention*. 2022
- Setting on the **sphere**:
B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. *A mathematical perspective on transformers*. 2025
- **Continuum** attention operator:
E. Calvello, N. B. Kovachki, M. E. Levine, and A. M. Stuart. "Continuum Attention for Neural Operators". In: *Journal of Machine Learning Research* 26.300 (2025), pp. 1–52

Overview of the transformer architecture

Embedding and encoding



Deutsche
Forschungsgemeinschaft

We consider the architecture from [Vas+17].

What is the day
after Monday?

Embedding

$$X \in \mathbb{R}^{N \times d}$$

Encoder

$$Z \in \mathbb{R}^{N \times d}$$

Overview of the transformer architecture

Embedding and encoding



We consider the architecture from [Vas+17].

What is the day
after Monday?

Embedding

$$X \in \mathbb{R}^{N \times d}$$

Encoder

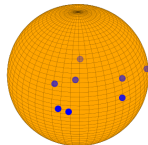
$$Z \in \mathbb{R}^{N \times d}$$

Details on embedding:

What is the day
after Monday?

[what][is][the][day]
[after][Mon][day][?]

Embedding +
positional Encoding



$$X \in \mathbb{R}^{N \times d}$$

Note: The embedding acts *locally* on each token.

Overview of the transformer architecture

Embedding and encoding



We consider the architecture from [Vas+17].

What is the day
after Monday?

Embedding

$$X \in \mathbb{R}^{N \times d}$$

Encoder

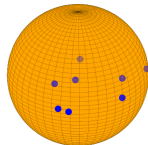
$$Z \in \mathbb{R}^{N \times d}$$

Details on embedding:

What is the day
after Monday?

[what][is][the][day]
[after][Mon][day][?]

Embedding +
positional Encoding



$$X \in \mathbb{R}^{N \times d}$$

Note: The embedding acts *locally* on each token.

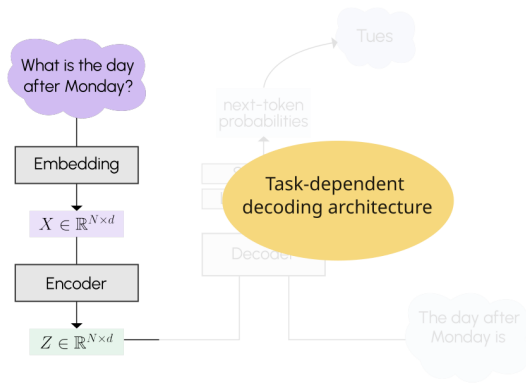
What is the role of the encoder?

Overview of the transformer architecture

Encoder-only perspective



We consider the architecture from [Vas+17].



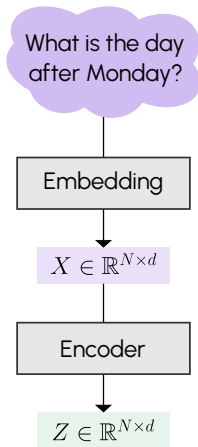
Overview of the transformer architecture

Role of the encoder



Deutsche
Forschungsgemeinschaft

We consider the architecture from [Vas+17].



Note: The embedding acts *locally* on each token.

What is the role of the encoder?

- Generate a *suitable input representation* for the upcoming task.

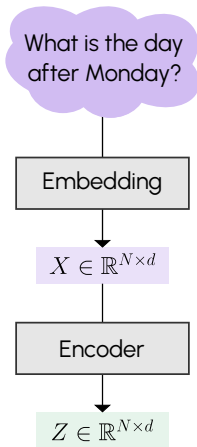
Overview of the transformer architecture

Role of the encoder



Deutsche
Forschungsgemeinschaft

We consider the architecture from [Vas+17].



Note: The embedding acts *locally* on each token.

What is the role of the encoder?

- Generate a *suitable input representation* for the upcoming task.
- Heuristically: Let embedded vectors interact with each other to *localize meaning*.

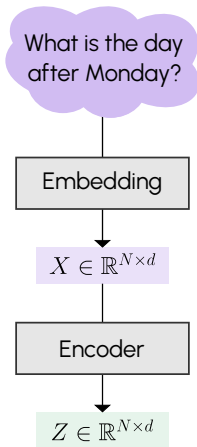
Overview of the transformer architecture

Role of the encoder



Deutsche
Forschungsgemeinschaft

We consider the architecture from [Vas+17].



Note: The embedding acts *locally* on each token.

What is the role of the encoder?

- Generate a *suitable input representation* for the upcoming task.
- Heuristically: Let embedded vectors interact with each other to *localize meaning*.
 $\rightsquigarrow$ Self-attention layers.

Self-Attention Layers

Self-Attention Layers

Discrete architecture



Deutsche
Forschungsgemeinschaft

- Input $X = [X_i]_{i=1}^N \in \mathbb{R}^{N \times d}$,

Self-Attention Layers

Discrete architecture



Deutsche
Forschungsgemeinschaft

- Input $X = [X_i]_{i=1}^N \in \mathbb{R}^{N \times d}$,
- Parameters $D, V \in \mathbb{R}^{d \times d}$

Self-Attention Layers

Discrete architecture



- Input $X = [X_i]_{i=1}^N \in \mathbb{R}^{N \times d}$,
- Parameters $D, V \in \mathbb{R}^{d \times d}$
- Discrete layer $\mathcal{L}_{D,V} : \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^{N \times d}$,

$$\mathcal{L}_{D,V}(X)_i = X_i + V M_D(X)_i, \quad \text{with} \quad M_D(X)_i = \sum_{j=1}^N \frac{\exp(X_i \cdot DX_j)}{\sum_{k=1}^N \exp(X_i \cdot DX_k)} X_j,$$

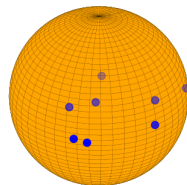
for $i = 1, \dots, N$.

Self-Attention Layers

Discrete encoder architecture with normalization



- **Mimic layer normalization:** $X = [X_i]_{i=1}^N \in \mathcal{S}^N$, with sphere $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\| = 1\}$.

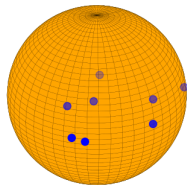


Self-Attention Layers

Discrete encoder architecture with normalization



- Mimic layer normalization: $X = [X_i]_{i=1}^N \in \mathcal{S}^N$, with sphere $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\| = 1\}$.
- Projection $P(x) = \frac{x}{\|x\|}$ for $x \in \mathbb{R}^d \setminus \{0\}$.



Self-Attention Layers

Discrete encoder architecture with normalization

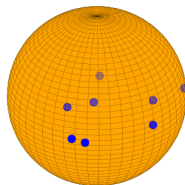


Deutsche
Forschungsgemeinschaft

- **Mimic layer normalization:** $X = [X_i]_{i=1}^N \in \mathcal{S}^N$, with sphere $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\| = 1\}$.
- Projection $P(x) = \frac{x}{\|x\|}$ for $x \in \mathbb{R}^d \setminus \{0\}$.
- Normalized layer $\mathcal{L}_{D,V} : \mathcal{S}^N \rightarrow \mathcal{S}^N$ ([Ges+25]),

$$\mathcal{L}_{D,V}(X)_i = P(X_i + V M_D(X)_i)$$

for $i = 1, \dots, N$.



Self-Attention Layers

Discrete encoder architecture with normalization

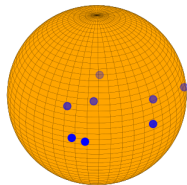


- **Mimic layer normalization:** $X = [X_i]_{i=1}^N \in \mathcal{S}^N$, with sphere $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\| = 1\}$.
- Projection $P(x) = \frac{x}{\|x\|}$ for $x \in \mathbb{R}^d \setminus \{0\}$.
- Normalized layer $\mathcal{L}_{D,V} : \mathcal{S}^N \rightarrow \mathcal{S}^N$ ([Ges+25]),

$$\mathcal{L}_{D,V}(X)_i = P(X_i + V M_D(X)_i)$$

for $i = 1, \dots, N$.

- **Assume** $\|V\| < 1$ for well-definedness.



Self-Attention Layers

Fixed points in finite dimensions



Deutsche
Forschungsgemeinschaft

- If $\mathcal{L}_{D,V}(X) = X$ we call X a fixed point.

Self-Attention Layers

Fixed points in finite dimensions



Deutsche
Forschungsgemeinschaft

- If $\mathcal{L}_{D,V}(X) = X$ we call X a fixed point.
- Orthogonal complement of P at $x \in \mathcal{S}$ for $z \in \mathbb{R}^d$:

$$P_x^\perp(z) = z - (z \cdot x)x.$$

Self-Attention Layers

Fixed points in finite dimensions



- If $\mathcal{L}_{D,V}(X) = X$ we call X a fixed point.
- Orthogonal complement of P at $x \in \mathcal{S}$ for $z \in \mathbb{R}^d$:

$$P_x^\perp(z) = z - (z \cdot x)x.$$

- Recall $\|V\| < 1$. $X = [X_i]_{i=1}^N$ is a fixed point if and only if

$$P_{X_i}^\perp(V M_D(X)_i) = 0$$

for all $i = 1, \dots, N$.

Self-Attention Layers

Adaptation for measures



Deutsche
Forschungsgemeinschaft

- Consider probability measure $\mu \in \mathcal{P}(\mathcal{S})$, e.g.,

$$\mu = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}.$$

Self-Attention Layers

Adaptation for measures



Deutsche
Forschungsgemeinschaft

- Consider probability measure $\mu \in \mathcal{P}(\mathcal{S})$, e.g.,

$$\mu = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}.$$

- Infinite-dimensional layer $\mathcal{L}_{D,V} : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{S})$,

$$\mathcal{L}_{D,V}(\mu) = (P \circ (\text{Id} + V M_D[\mu]))_{\#} \mu,$$

with $M_D[\mu] : \mathcal{S} \rightarrow \mathbb{R}^d$,

$$M_D[\mu](x) = \int_{\mathcal{S}} \frac{\exp(x \cdot Dy)}{\int_{\mathcal{S}} \exp(x \cdot Dz) \, d\mu(z)} y \, d\mu(y).$$

Self-Attention Layers

Fixed points in infinite dimensions



Deutsche
Forschungsgemeinschaft

- If $\mathcal{L}_{D,V}(\mu) = \mu$ we call μ a fixed point.

Self-Attention Layers

Fixed points in infinite dimensions



Deutsche
Forschungsgemeinschaft

- If $\mathcal{L}_{D,V}(\mu) = \mu$ we call μ a fixed point.
- If for μ -almost any $x \in \mathcal{S}$ it holds that

$$P_x^\perp (V M_D[\mu](x)) = 0,$$

then μ is a fixed point.

Evolution of Measures

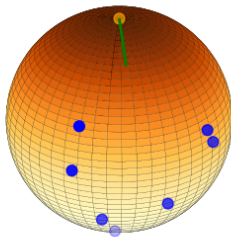
Evolution of measures

Motivation



Deutsche
Forschungsgemeinschaft

Compose multiple self-attention layers with shared weights (like ALBERT [Lan+19])



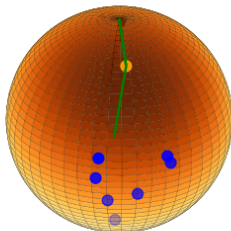
Evolution of measures

Motivation



Deutsche
Forschungsgemeinschaft

Compose multiple self-attention layers with shared weights (like ALBERT [Lan+19])



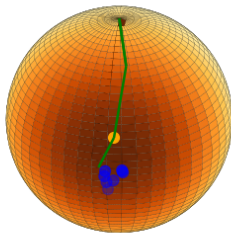
Evolution of measures

Motivation



Deutsche
Forschungsgemeinschaft

Compose multiple self-attention layers with shared weights (like ALBERT [Lan+19])



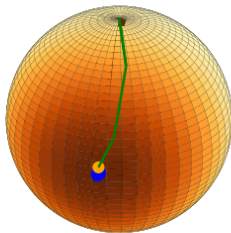
Evolution of measures

Motivation



Deutsche
Forschungsgemeinschaft

Compose multiple self-attention layers with shared weights (like ALBERT [Lan+19])



Evolution of Measures

Continuity equation [Ges+25]



Deutsche
Forschungsgemeinschaft

For a probability measure $\mu \in \mathcal{P}(\mathcal{S})$ define the **velocity field**

$$\mathbf{V}_{D,V}[\mu](x) := P_x^\perp(V M_D[\mu](x))$$

Evolution of Measures

Continuity equation [Ges+25]



Deutsche
Forschungsgemeinschaft

For a probability measure $\mu \in \mathcal{P}(\mathcal{S})$ define the **velocity field**

$$\mathbf{V}_{D,V}[\mu](x) := P_x^\perp(V M_D[\mu](x))$$

and study the **continuity equation**

$$\begin{aligned} \partial_t \mu + \operatorname{div}(\mathbf{V}_{D,V}[\mu] \mu) &= 0 && \text{on } [0, T] \times \mathcal{S}, \\ \mu|_{t=0} &= \mu_0 && \text{on } \mathcal{S}. \end{aligned}$$

For D symmetric and $V = \pm D$: Stationary states are critical points of the interaction energy

$$\mathcal{E}(\mu) := \int_{\mathcal{S}} \int_{\mathcal{S}} \exp(x \cdot Dy) \, d\mu(x) d\mu(y).$$

For D symmetric and $V = \pm D$: Stationary states are critical points of the interaction energy

$$\mathcal{E}(\mu) := \int_{\mathcal{S}} \int_{\mathcal{S}} \exp(x \cdot Dy) \, d\mu(x) d\mu(y).$$

$\rightsquigarrow$ Self-attention dynamics as gradient flow [Ges+24; Bur+25]

For D symmetric and $V = \pm D$: Stationary states are critical points of the interaction energy

$$\mathcal{E}(\mu) := \int_{\mathcal{S}} \int_{\mathcal{S}} \exp(x \cdot Dy) \, d\mu(x) d\mu(y).$$

$\rightsquigarrow$ Self-attention dynamics as gradient flow [Ges+24; Bur+25]

Evolution of Measures

Interaction Energy



DFG Deutsche
Forschungsgemeinschaft

For D symmetric and $V = \pm D$: Stationary states are critical point energy

$$\mathcal{E}(\mu) := \int_{\mathcal{S}} \int_{\mathcal{S}} \exp(x \cdot Dy) \, d\mu(x) d\mu(y).$$

$\rightsquigarrow$ Self-attention dynamics as gradient flow [Ges+24; Bur+25]



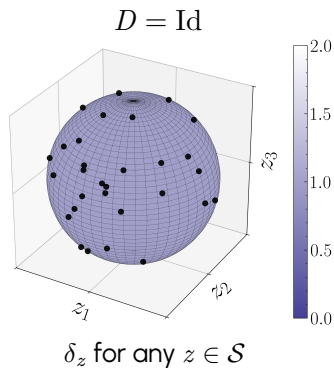
lukas.weigand@desy.de

Stationary States

Attraction – Maximizers

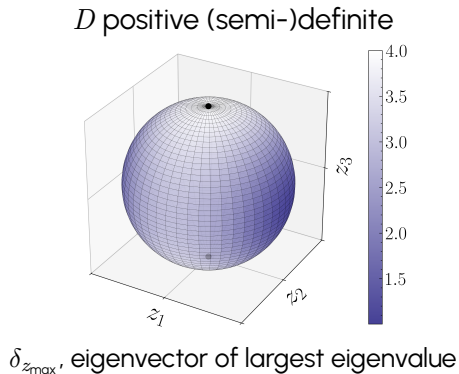
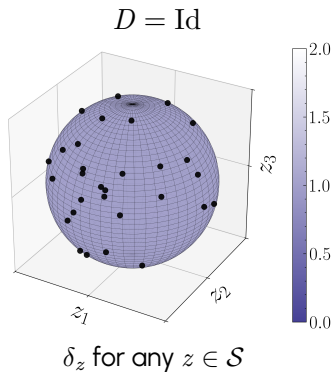


Deutsche
Forschungsgemeinschaft



Stationary States

Attraction – Maximizers



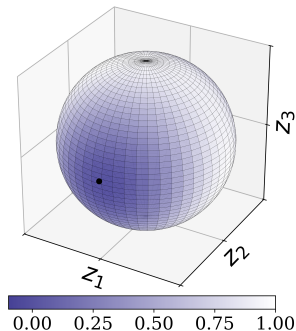
Interaction Energy

Attraction – Minimizers



Deutsche
Forschungsgemeinschaft

D indefinite or negative (semi-)definite



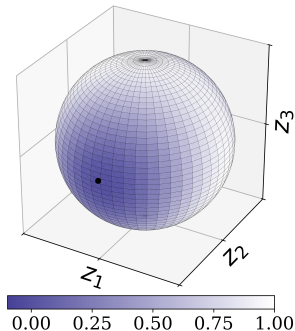
$\delta_{z_{\min}}$, eigenvector of smallest eigenvalue

Interaction Energy

Attraction – Minimizers

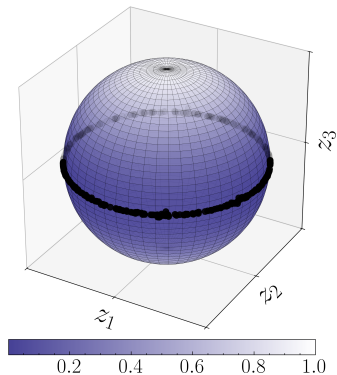


D indefinite or negative (semi-)definite



$\delta_{z_{\min}}$, eigenvector of smallest eigenvalue

D positive semi-definite



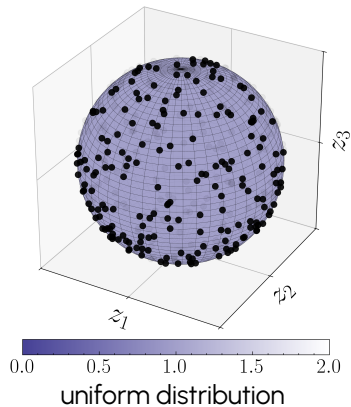
any measure concentrated on $\mathcal{N}(D)$

Interaction Energy

Repulsion – Minimizers



$$D = \text{Id}$$



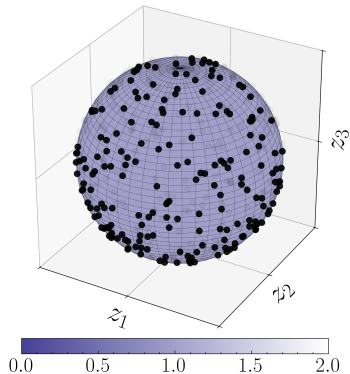
Interaction Energy

Repulsion – Minimizers



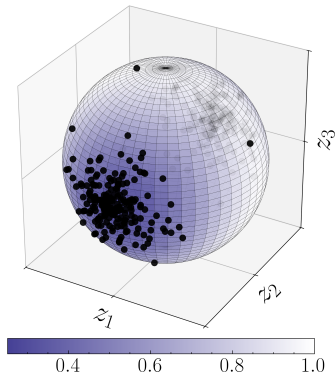
Deutsche
Forschungsgemeinschaft

$$D = \text{Id}$$



uniform distribution

D positive definite (ongoing work!)



$\mathcal{E}$ strictly convex

$\rightsquigarrow$ uniqueness, symmetry [BMV22]

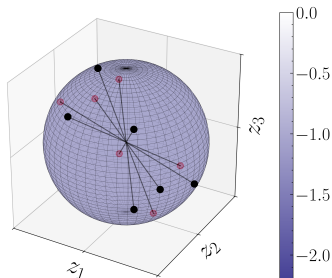
Interaction Energy

Repulsion – Maximizers



Deutsche
Forschungsgemeinschaft

$$D = -\text{Id}$$



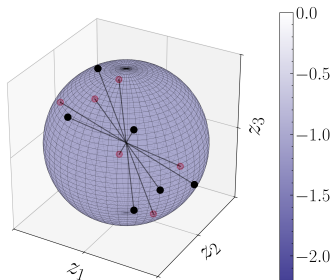
$$1/2(\delta_z + \delta_{-z}) \text{ for any } z \in \mathcal{S}$$

Interaction Energy

Repulsion – Maximizers

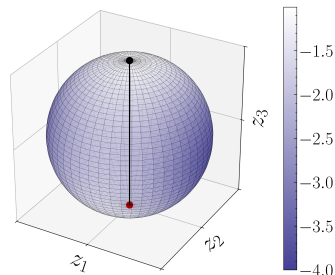


$$D = -\text{Id}$$



$1/2(\delta_z + \delta_{-z})$ for any $z \in \mathcal{S}$

D negative (semi-)definite



$1/2(\delta_{z_{\min}} + \delta_{-z_{\min}})$, eigenvector of smallest eigenvalue

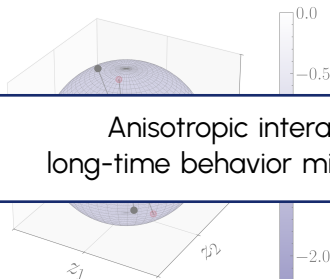
Interaction Energy

Repulsion – Maximizers

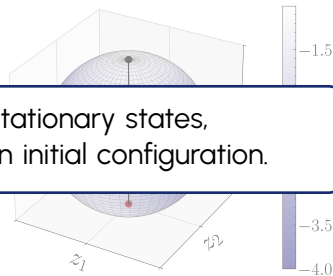


$$D = -\text{Id}$$

D negative (semi-)definite



$$1/2(\delta_z + \delta_{-z}) \text{ for any } z \in \mathcal{S}$$



$$1/2(\delta_{z_{\min}} + \delta_{-z_{\min}}), \text{ eigenvector of smallest eigenvalue}$$

Conclusion and Outlook

In this talk:

- Transformer architecture with self-attention layers and layer normalization [[Vas+17](#)]
- Adaptation to probability measures [[Cal+25](#)] and time-continuous evolution of measures [[San+22](#); [Ges+25](#)]
- Long-time behavior and energy optimization: Optimizers depend on interaction matrix [[Bur+25](#)]

In this talk:

- Transformer architecture with self-attention layers and layer normalization [Vas+17]
- Adaptation to probability measures [Cal+25] and time-continuous evolution of measures [San+22; Ges+25]
- Long-time behavior and energy optimization: Optimizers depend on interaction matrix [Bur+25]

A **big open question**: Practical implications?

In this talk:

- Transformer architecture with self-attention layers and layer normalization [[Vas+17](#)]
- Adaptation to probability measures [[Cal+25](#)] and time-continuous evolution of measures [[San+22](#); [Ges+25](#)]
- Long-time behavior and energy optimization: Optimizers depend on interaction matrix [[Bur+25](#)]

A [big open question](#): Practical implications?

Future work in this regard:

- Meta-stability for anisotropic interaction (isotropic in [[Ges+24](#)])

In this talk:

- Transformer architecture with self-attention layers and layer normalization [[Vas+17](#)]
- Adaptation to probability measures [[Cal+25](#)] and time-continuous evolution of measures [[San+22](#); [Ges+25](#)]
- Long-time behavior and energy optimization: Optimizers depend on interaction matrix [[Bur+25](#)]

A [big open question](#): Practical implications?

Future work in this regard:

- Meta-stability for anisotropic interaction (isotropic in [[Ges+24](#)])
- Spectral proof techniques for $D = \text{Id}$: Transferable to general D ? ([[PS25](#); [SS26](#)], Anna Shalova, Uni Amsterdam)

Thank you for your attention!
Questions?

 **HELMHOLTZ
IMAGING**



DFG Deutsche
Forschungsgemeinschaft

- [HSG] B. Hoover, H. Strobelt, and S. Gehrmann. *exBERT: A visual analysis tool to explore learned representations in transformer models*. URL: <https://huggingface.co/spaces/exbert-project/exbert>.
- [Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention is all you need*. 2017.
- [San+22] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. *Sinkformers: Transformers with doubly stochastic attention*. 2022.
- [Ges+25] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. *A mathematical perspective on transformers*. 2025.
- [Lan+19] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. *Albert: A lite bert for self-supervised learning of language representations*. 2019.
- [Ges+24] B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet. *Dynamic metastability in the self-attention model*. 2024.
- [Bur+25] M. Burger, S. Kabri, Y. Korolev, T. Roith, and Weigand. *Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization*. 2025.
- [BMV22] D. Bilyk, R. W. Matzke, and O. Vlasniuk. *Positive definiteness and the Stolarsky invariance principle*. 2022.
- [Cal+25] E. Calvello, N. B. Kovachki, M. E. Levine, and A. M. Stuart. "Continuum Attention for Neural Operators". In: *Journal of Machine Learning Research* 26:300 (2025), pp. 1–52.
- [PS25] M. A. Peletier and A. Shalova. *Nonlinear diffusion limit of non-local interactions on a sphere*. 2025.
- [SS26] A. Shalova and A. Schlichting. *Solutions of stationary McKean–Vlasov equation on a high-dimensional sphere and other Riemannian manifolds*. 2026.