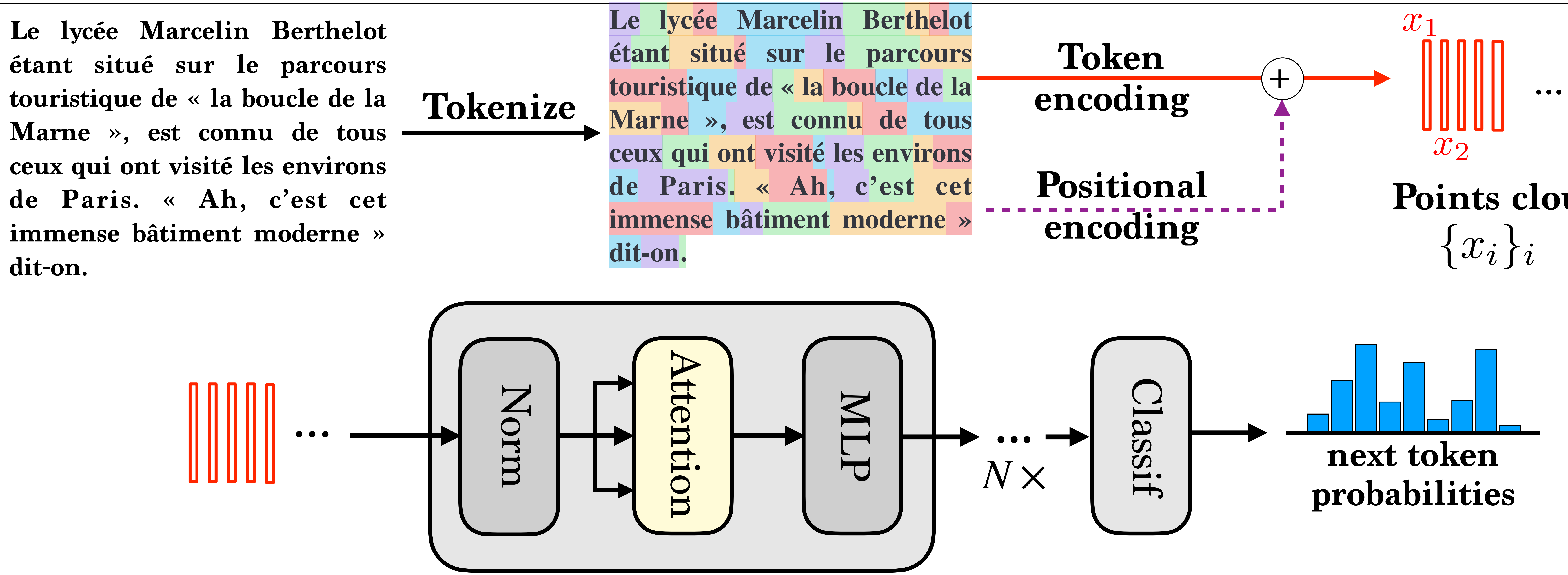
Transformers are Universal in Context Learners

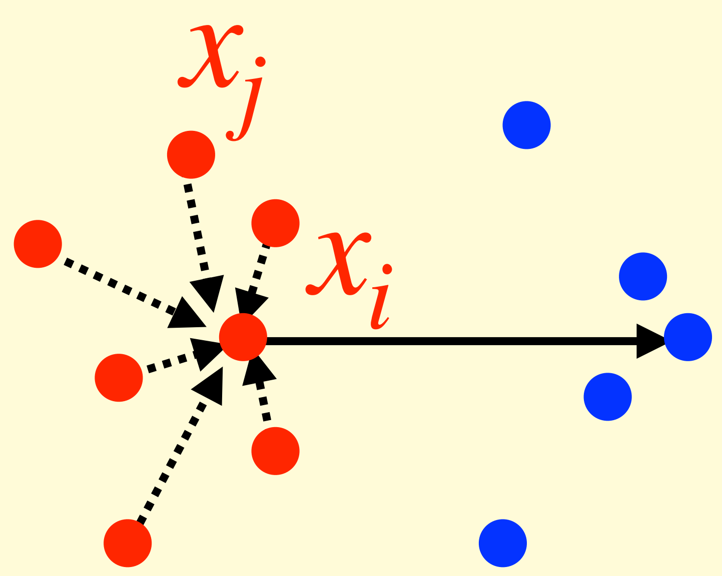
Gabriel Peyré

ENS
ÉCOLE NORMALE SUPÉRIEURE

Transformers and attention mechanism

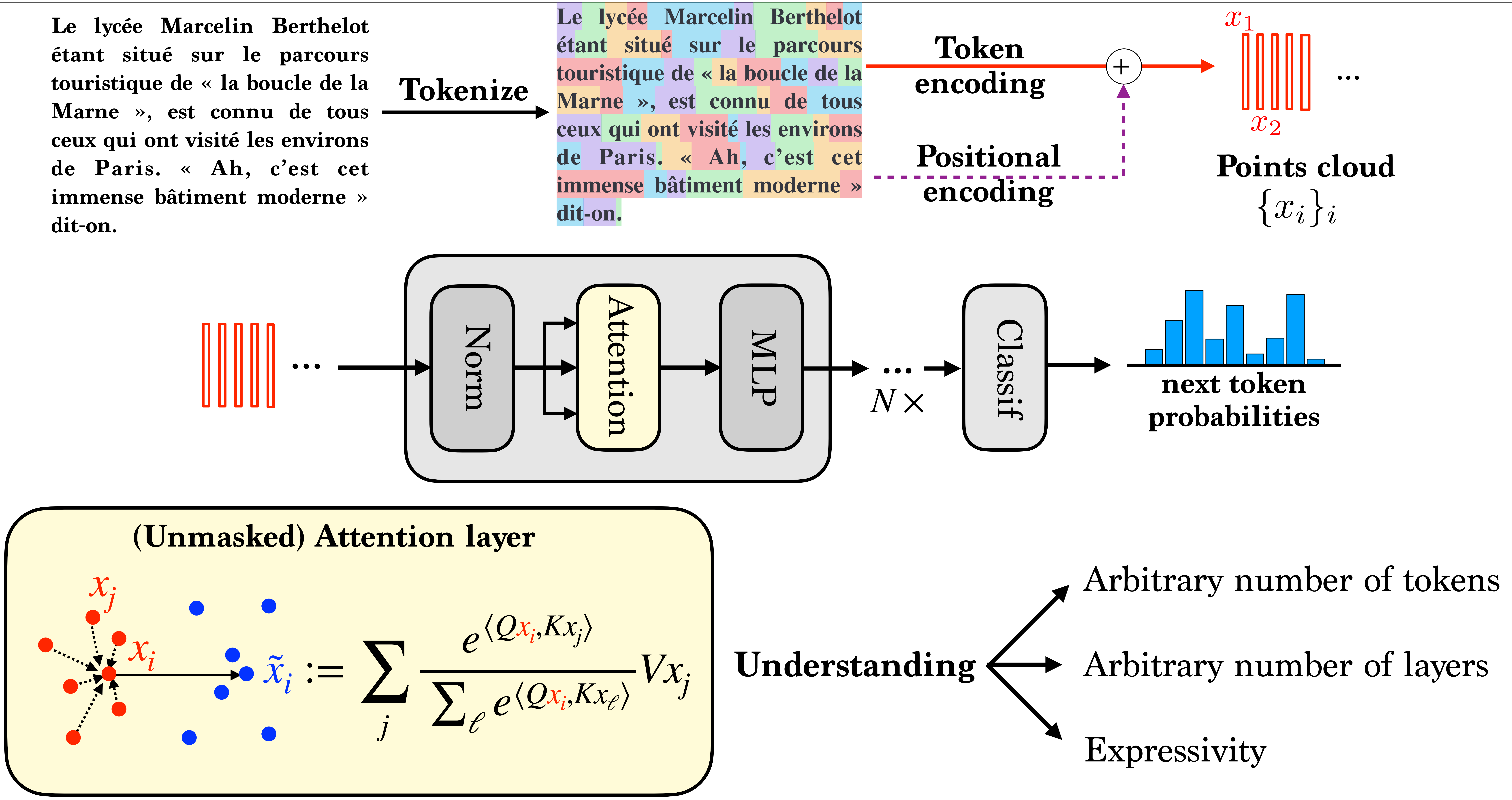


(Unmasked) Attention layer



$$\tilde{x}_i := \sum_j \frac{e^{\langle Qx_i, Kx_j \rangle}}{\sum_\ell e^{\langle Qx_i, Kx_\ell \rangle}} V x_j$$

Transformers and attention mechanism



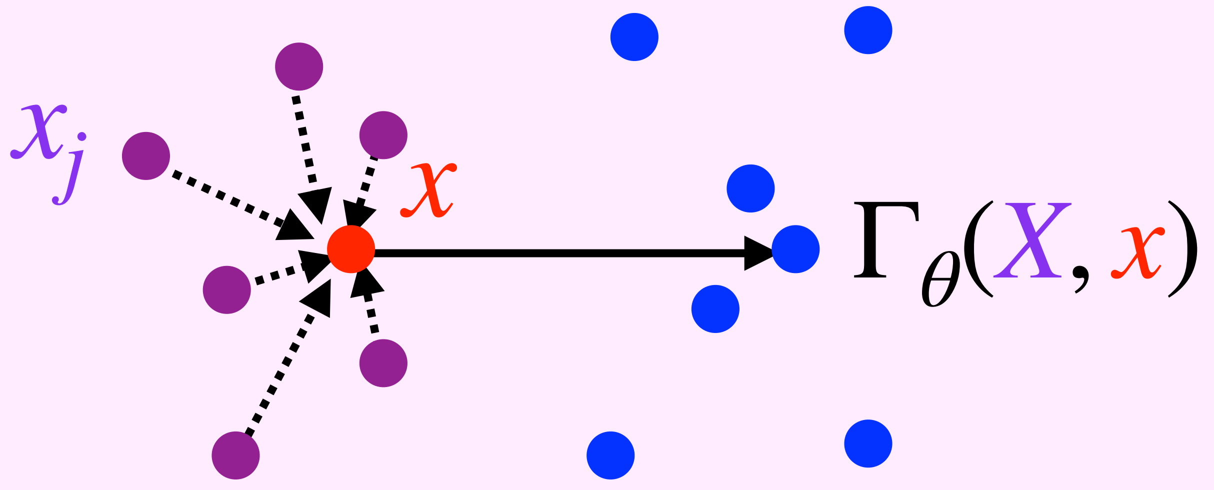

Arbitrary number of tokens
In Context Mappings over Measures
Arbitrary number of layers
Smoothness and PDE's
Expressivity
Universality

Attention as In-context Mapping

Point clouds: $X := \{x_i\}_{i=1}^n$

In-context mapping:
parameters $\theta := (Q, K, V)$

$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j \rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell \rangle}} V x_j$$

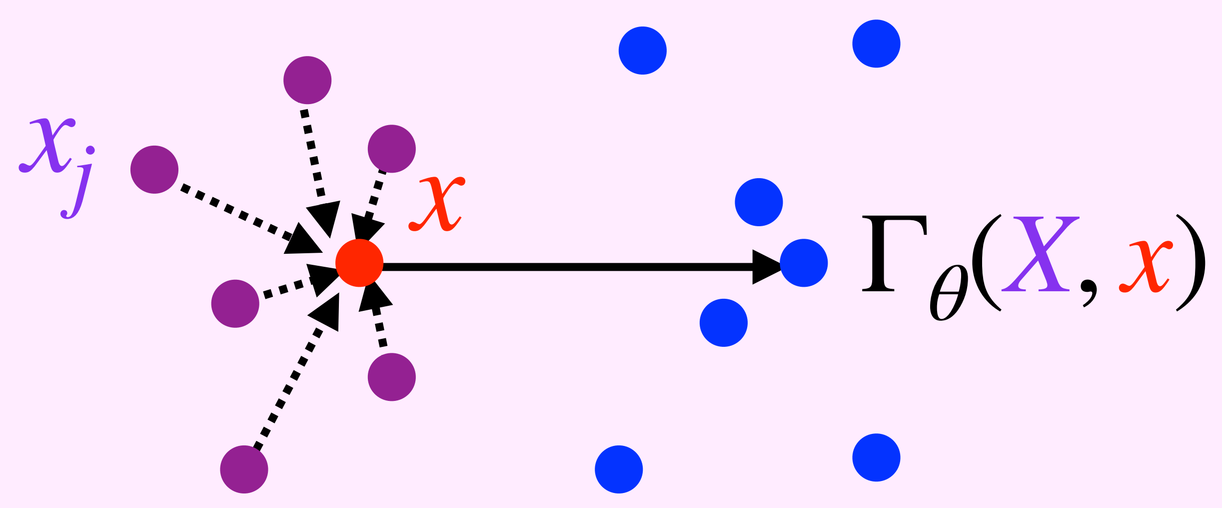


Attention as In-context Mapping

Point clouds: $X := \{x_i\}_{i=1}^n$

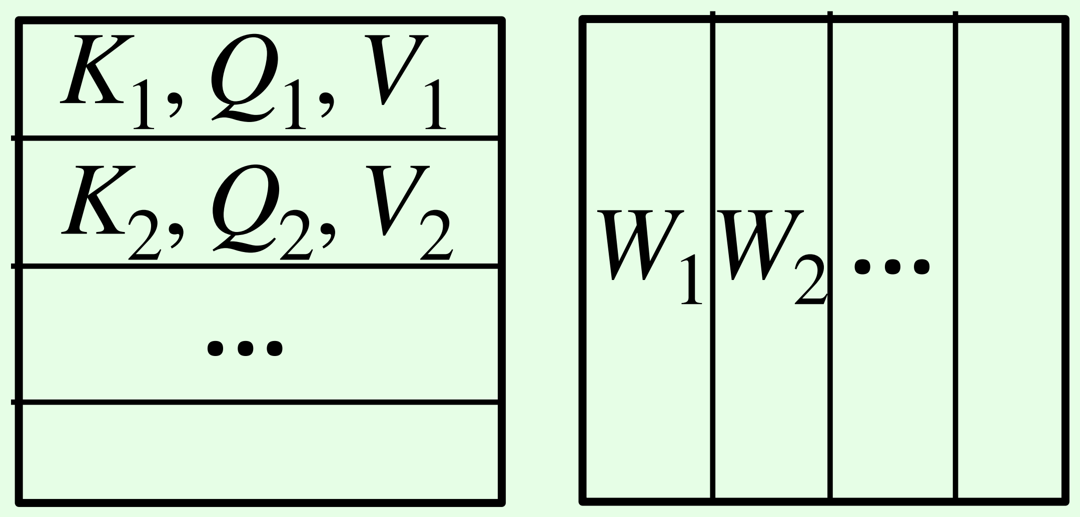
In-context mapping:
parameters $\theta := (Q, K, V)$

$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j \rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell \rangle}} V x_j$$



Single-head attention layer: $X \mapsto \left\{ \Gamma_\theta[X](x_i) \right\}_{i=1}^n$

Multi-head attention layer: $X \mapsto \left\{ \sum_{h=1}^H W_h \, \Gamma_{\theta_h}[X](x_i) \right\}_{i=1}^n$

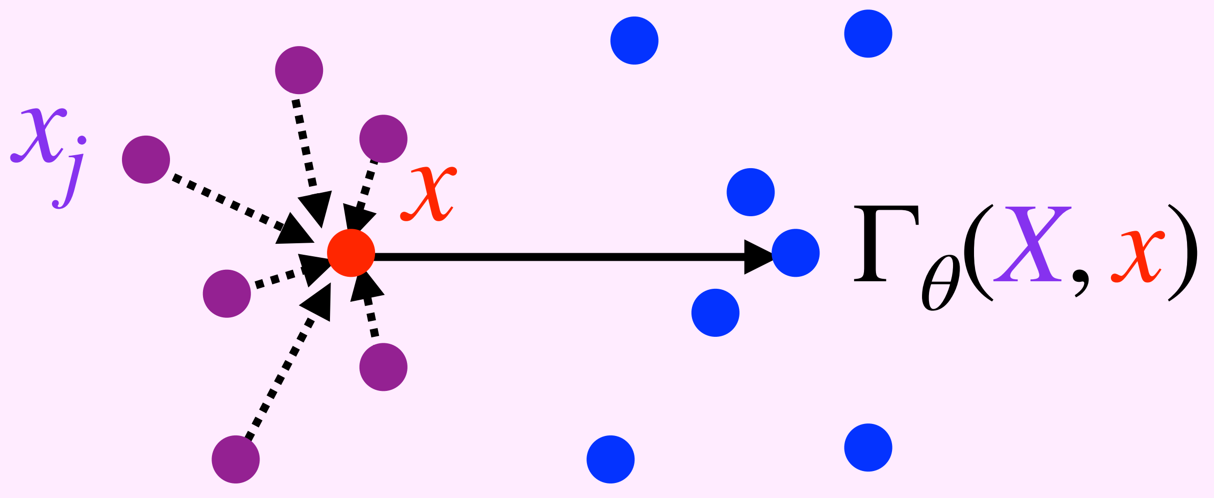


Attention as In-context Mapping

Point clouds: $X := \{x_i\}_{i=1}^n$

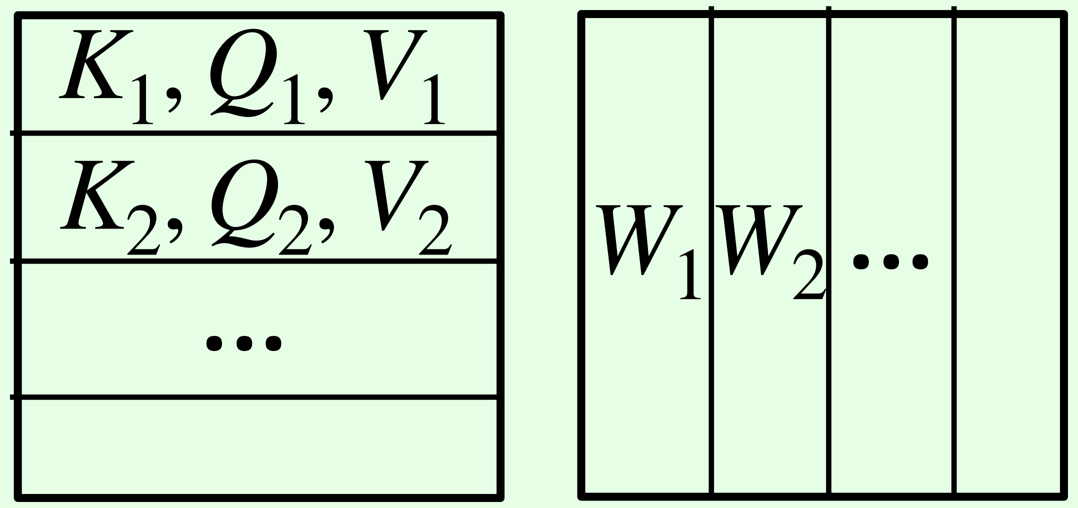
In-context mapping:
parameters $\theta := (Q, K, V)$

$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j \rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell \rangle}} V x_j$$



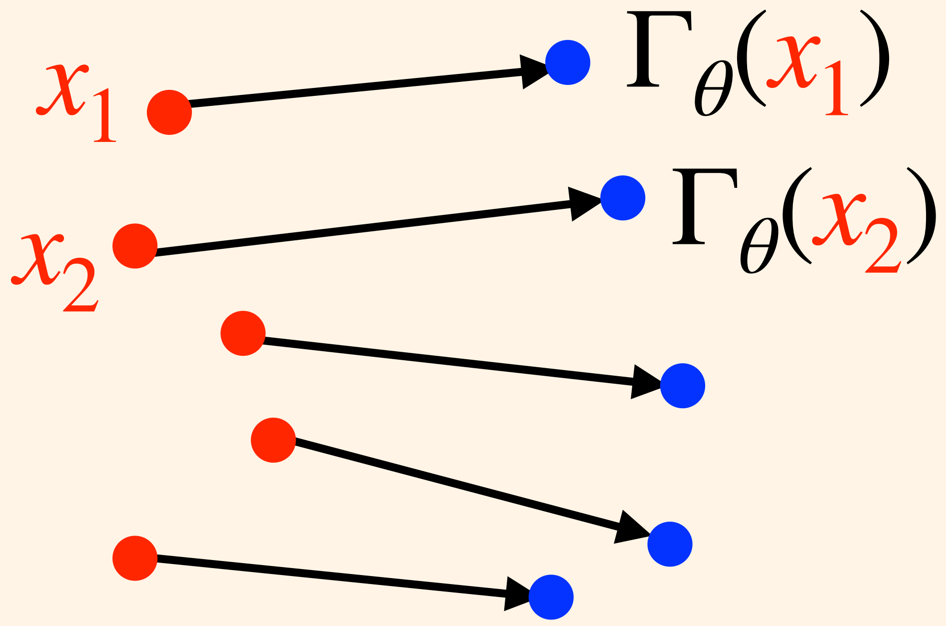
Single-head attention layer: $X \mapsto \left\{\Gamma_\theta[X](x_i)\right\}_{i=1}^n$

Multi-head attention layer: $X \mapsto \left\{\sum_{h=1}^H W_h \, \Gamma_{\theta_h}[X](x_i)\right\}_{i=1}^n$



Context-free layers: $X \mapsto \left\{\Gamma_\theta(x_i)\right\}_{i=1}^n$

Multi-layer perceptron: $\Gamma_\theta(x) := x + \theta_1 \mathrm{ReLu}(\theta_2 x)$

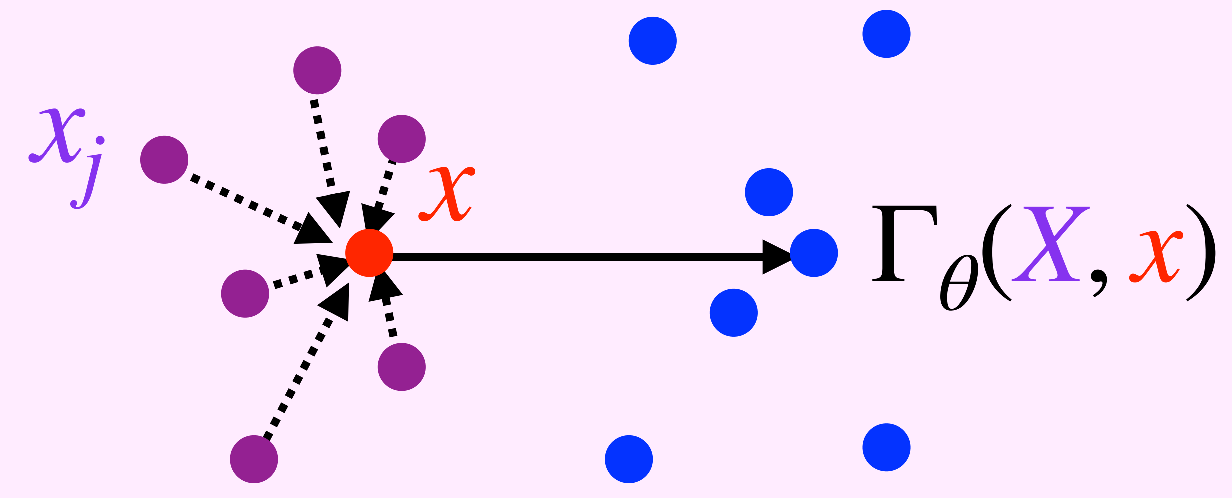


Attention as In-context Mapping

Point clouds: $X := \{x_i\}_{i=1}^n$

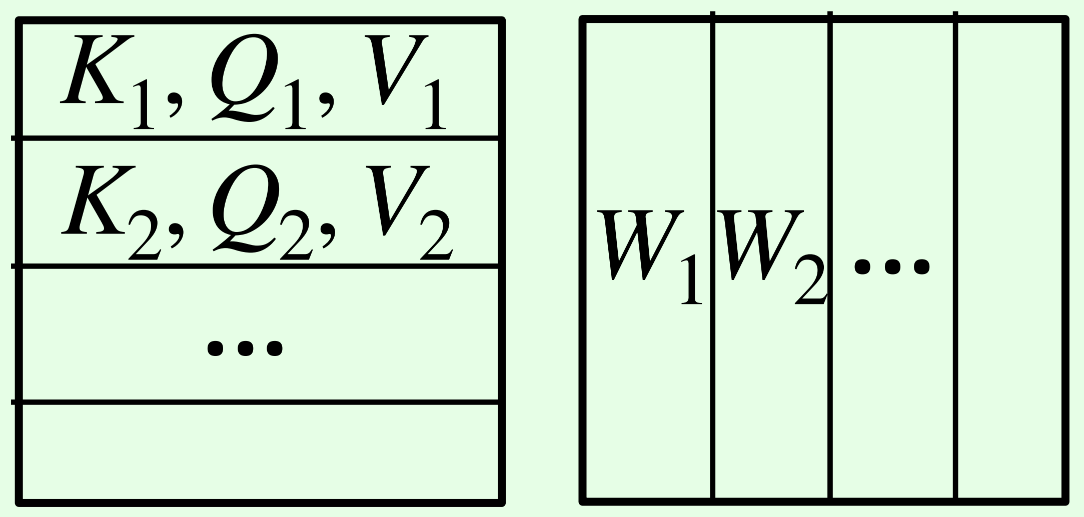
In-context mapping:
parameters $\theta := (Q, K, V)$

$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j \rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell \rangle}} V x_j$$



Single-head attention layer: $X \mapsto \left\{\Gamma_\theta[X](x_i)\right\}_{i=1}^n$

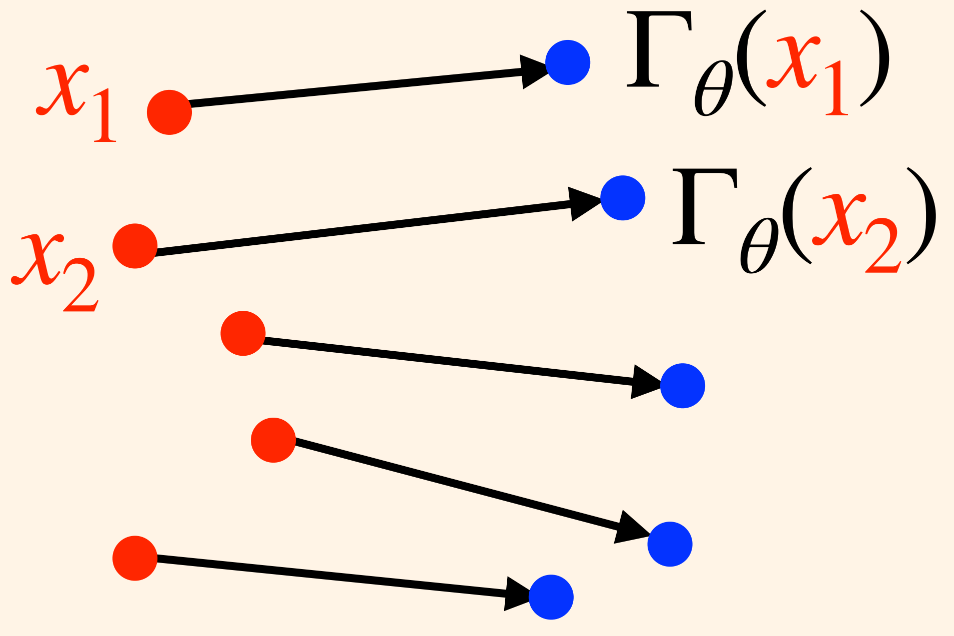
Multi-head attention layer: $X \mapsto \left\{\sum_{h=1}^H W_h \, \Gamma_{\theta_h}[X](x_i)\right\}_{i=1}^n$

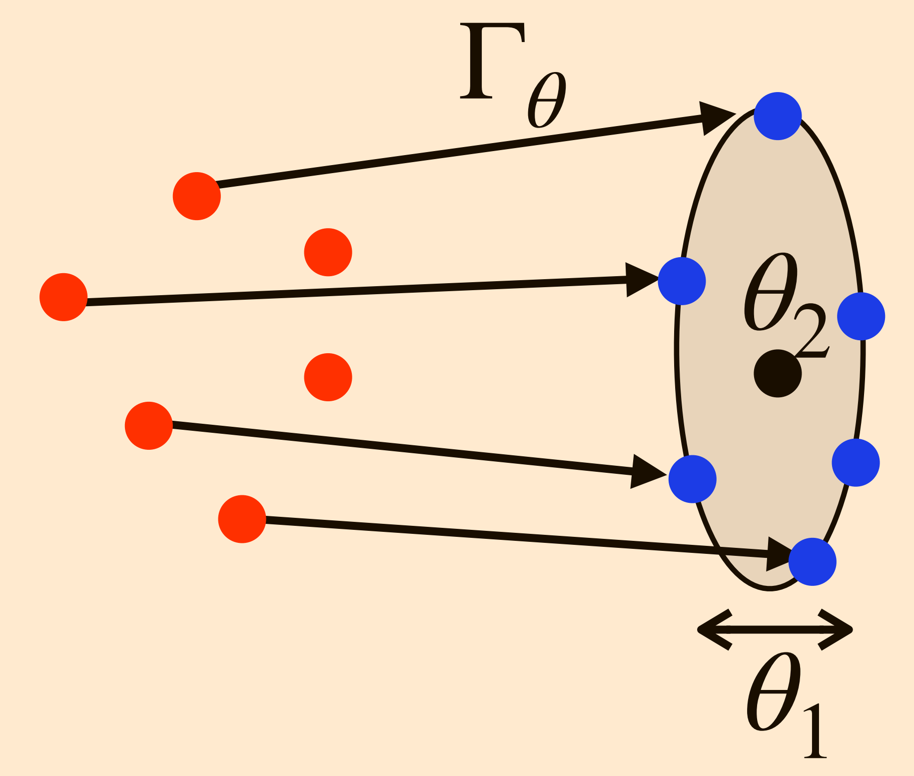


Context-free layers: $X \mapsto \left\{\Gamma_\theta(x_i)\right\}_{i=1}^n$

Multi-layer perceptron: $\Gamma_\theta(x) := x + \theta_1 \mathrm{ReLu}(\theta_2 x)$

Layer norm: $\Gamma_\theta(x) := \theta_1 \odot \frac{x}{\|x\|} + \theta_2$



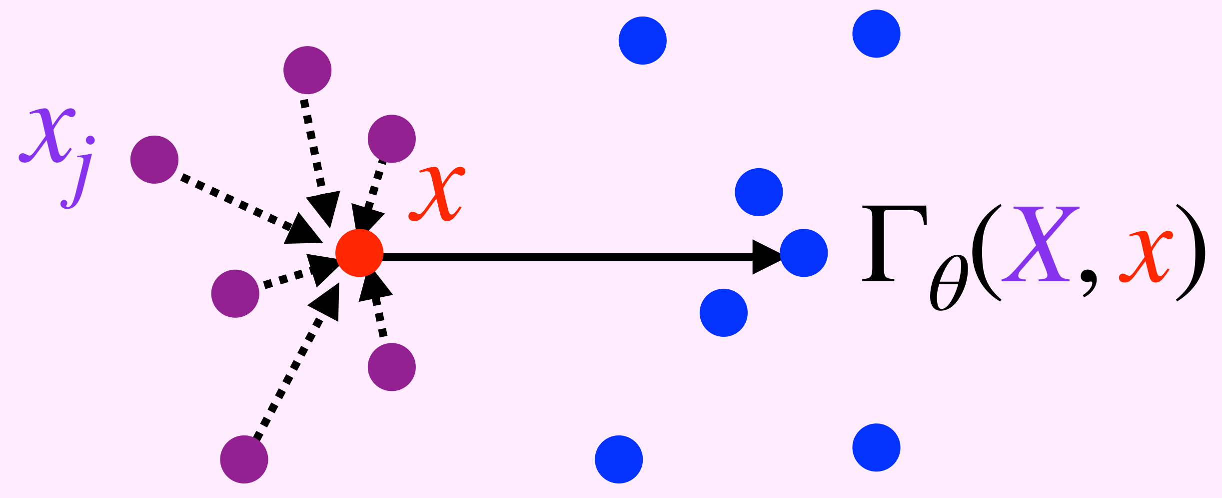


Attention as In-context Mapping

Point clouds: $X := \{x_i\}_{i=1}^n$

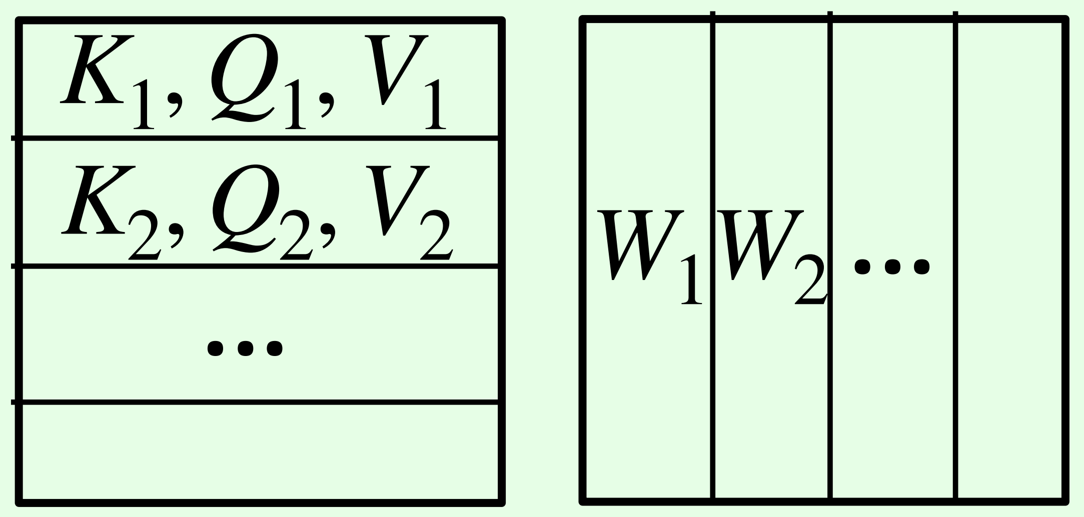
In-context mapping:
parameters $\theta := (Q, K, V)$

$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j \rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell \rangle}} V x_j$$

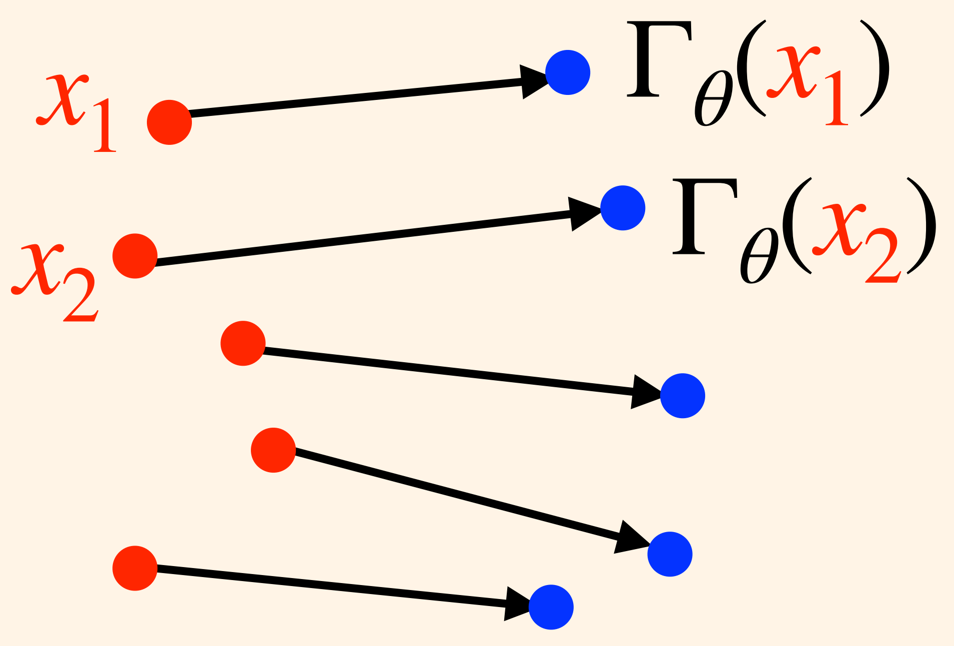


Single-head attention layer: $X \mapsto \left\{\Gamma_\theta[X](x_i)\right\}_{i=1}^n$

Multi-head attention layer: $X \mapsto \left\{\sum_{h=1}^H W_h \, \Gamma_{\theta_h}[X](x_i)\right\}_{i=1}^n$

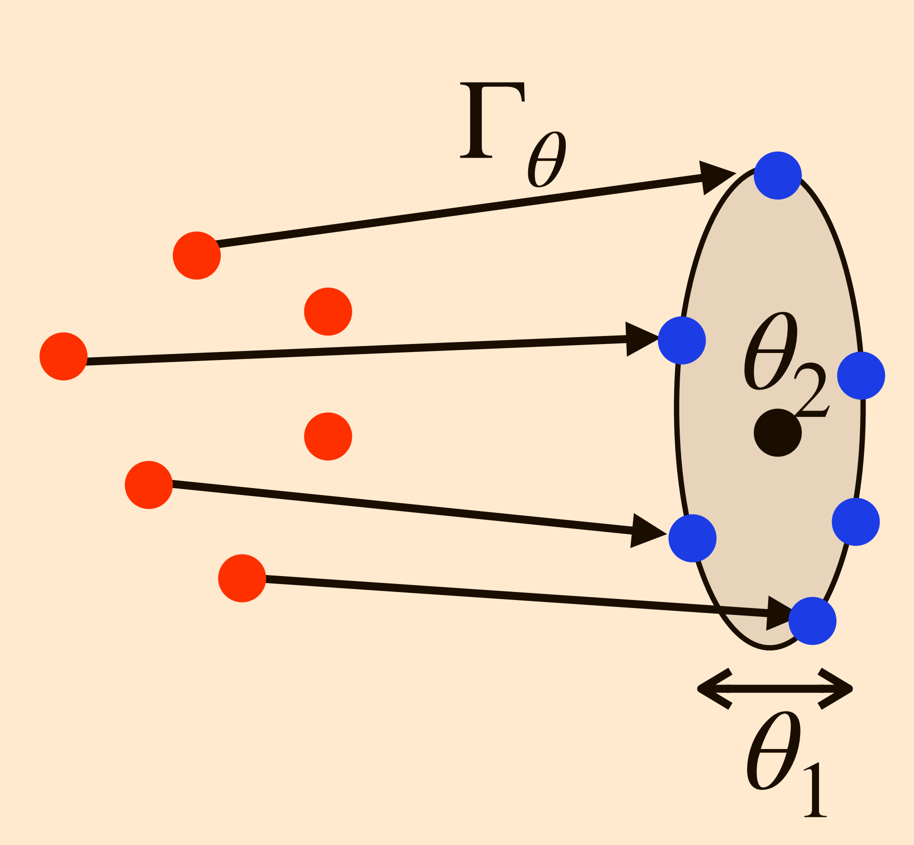


Context-free layers: $X \mapsto \left\{\Gamma_\theta(x_i)\right\}_{i=1}^n$



Multi-layer perceptron: $\Gamma_\theta(x) := x + \theta_1 \mathrm{ReLu}(\theta_2 x)$

Layer norm: $\Gamma_\theta(x) := \theta_1 \odot \frac{x}{\|x\|} + \theta_2$

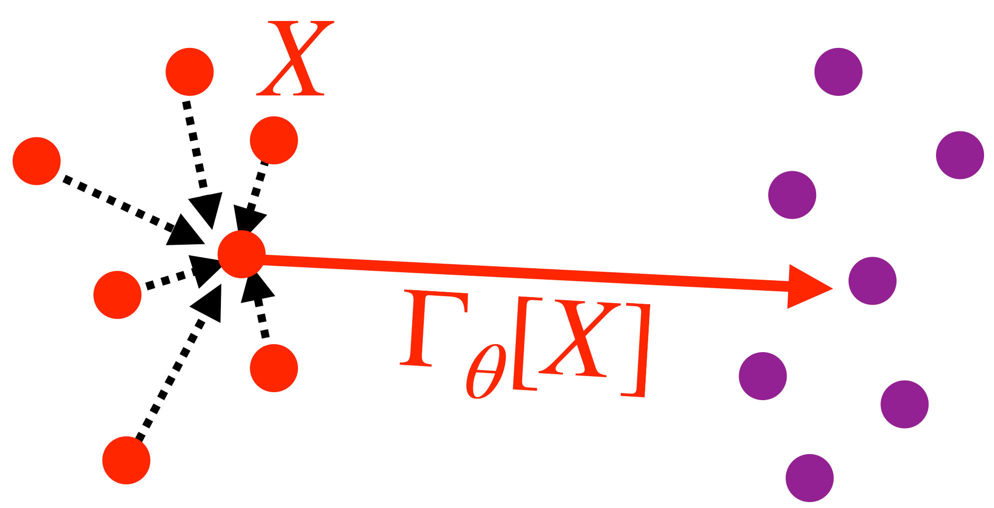


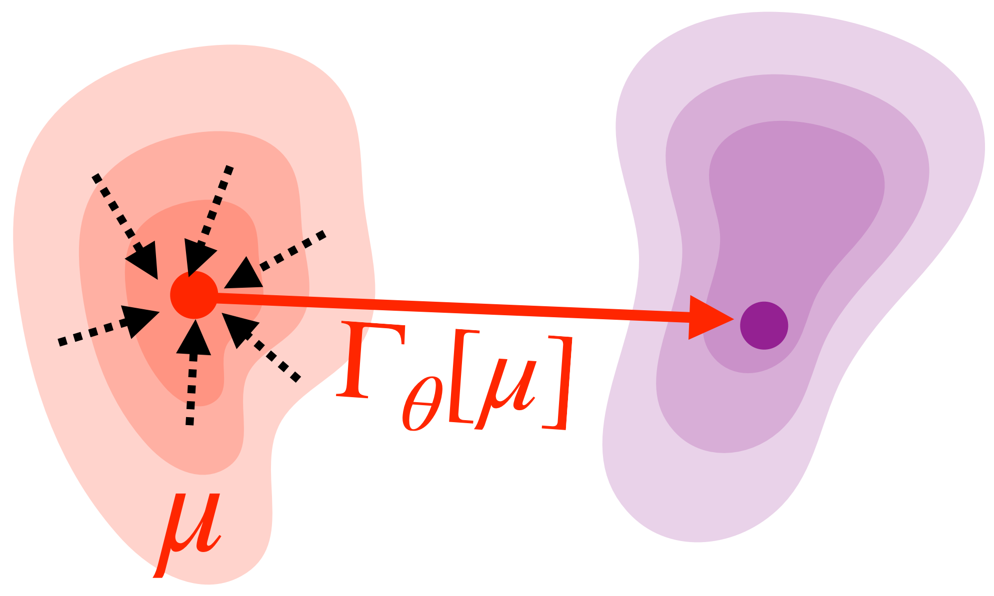
Transformer ≡ composition of in-context and context-free layers.

Attentions Operating over Measures

Number n of token is arbitrary. (Unmasked) attention is permutation invariant.

$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j \rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell \rangle}} V x_j \xrightarrow{\boxed{\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}}} \Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky \rangle}}{\int e^{\langle Qx, Ky' \rangle} \mathrm{d}\mu(y)} V y \, \mathrm{d}\mu(y)$$

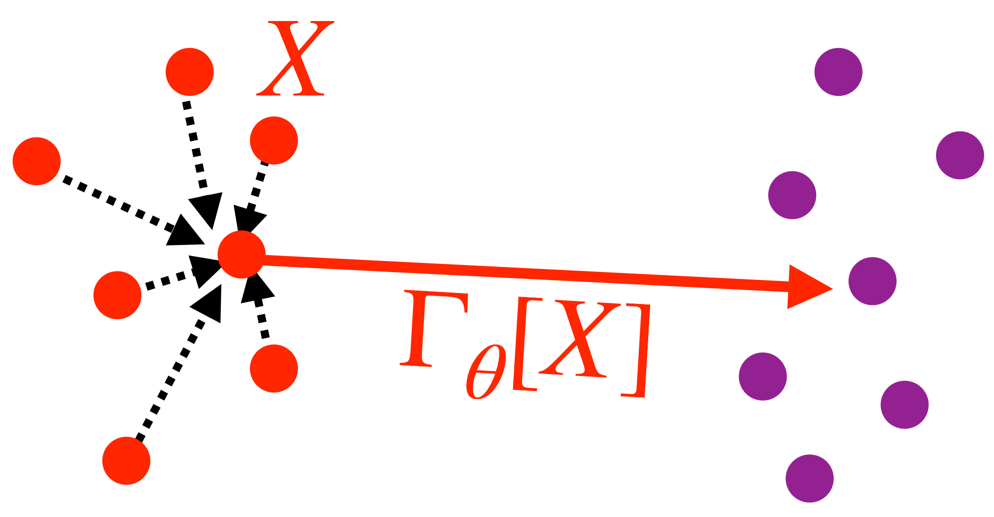


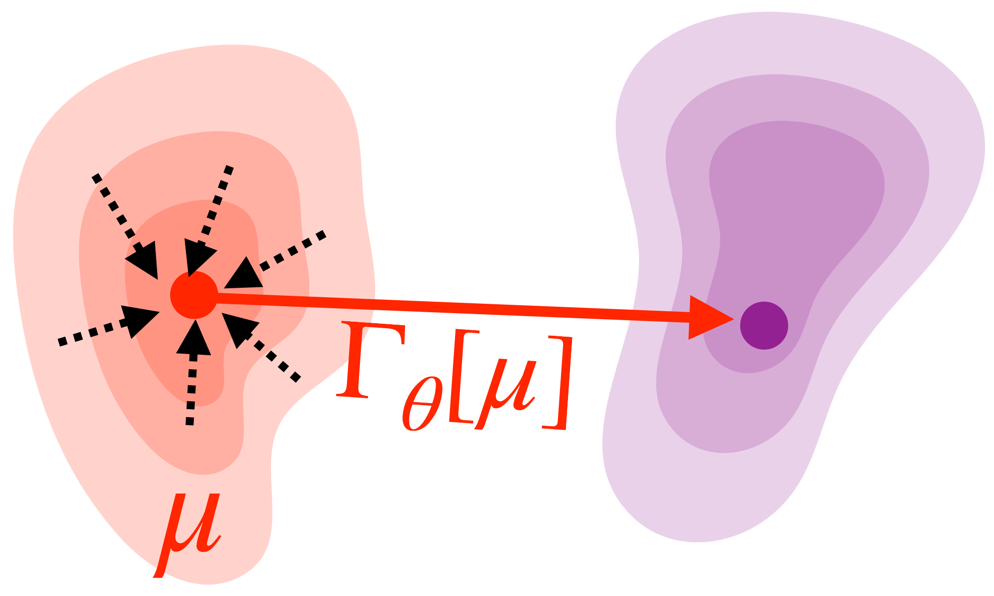


Attentions Operating over Measures

Number n of token is arbitrary. (Unmasked) attention is permutation invariant.

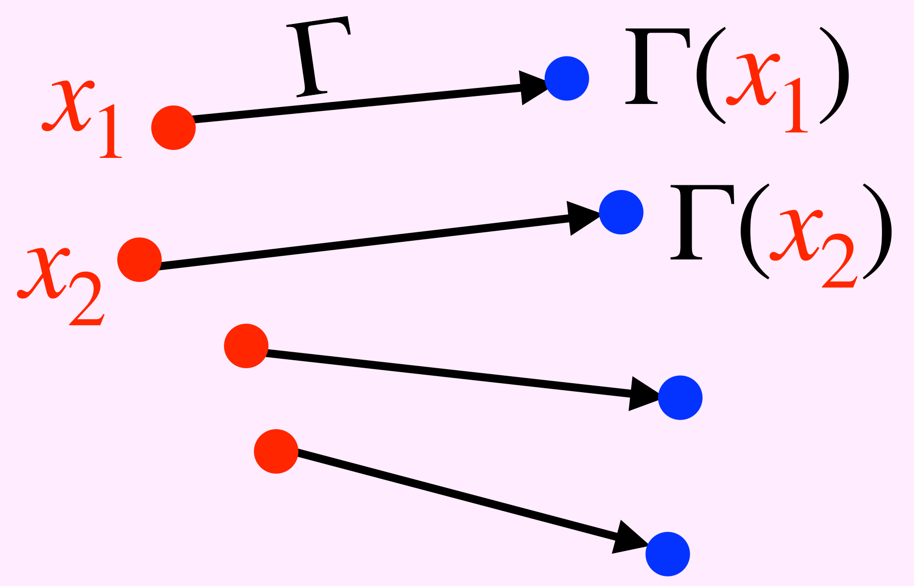
$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j\rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell\rangle}} V x_j \xrightarrow{\boxed{\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}}} \Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y)} V y \, \mathrm{d}\mu(y)$$





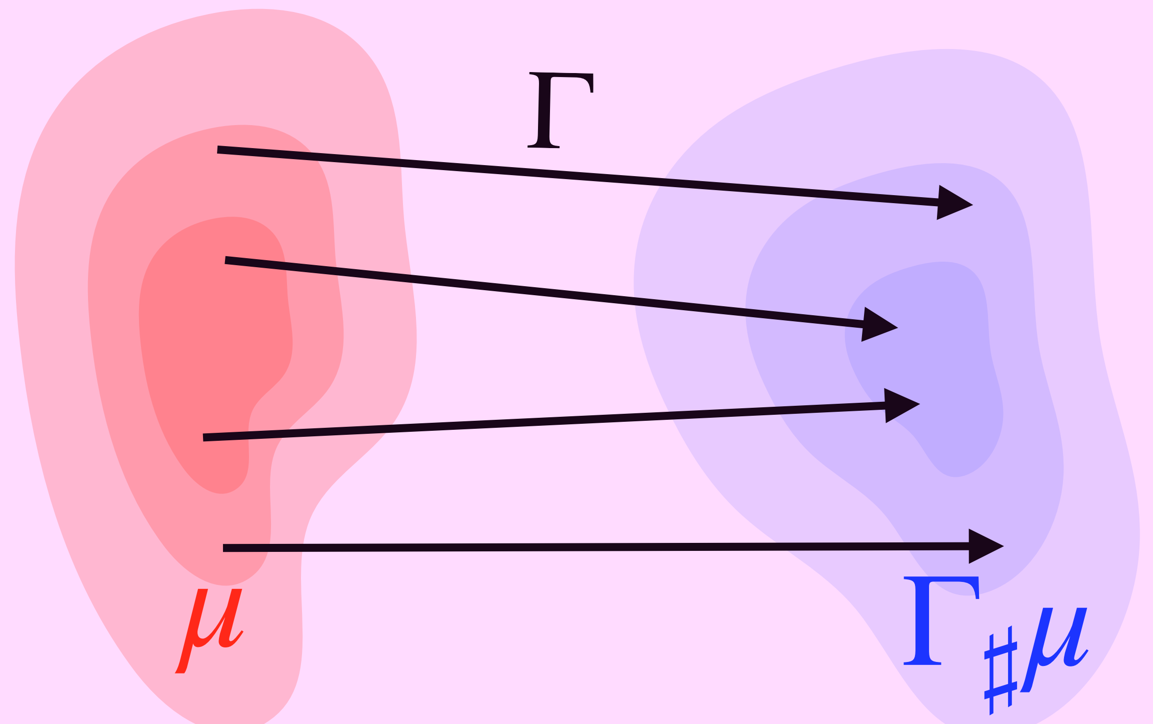
Push-forward

$$\Gamma_\sharp \sum_i \delta_{x_i} := \sum_i \delta_{\Gamma(x_i)}$$



$$\downarrow$$

$$(\Gamma_\sharp \mu)(B) := \mu(\Gamma^{-1}(B))$$



Attention layers

$$X \mapsto \{\Gamma_\theta[X](x_i)\}_{i=1}^n$$

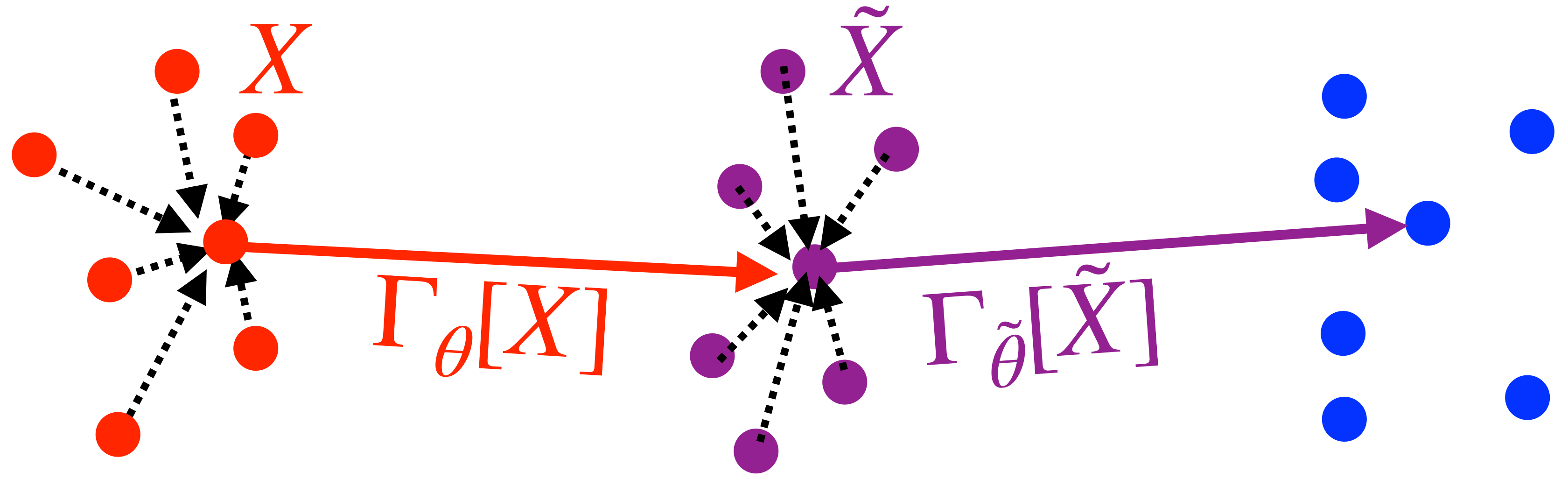
$$\downarrow$$

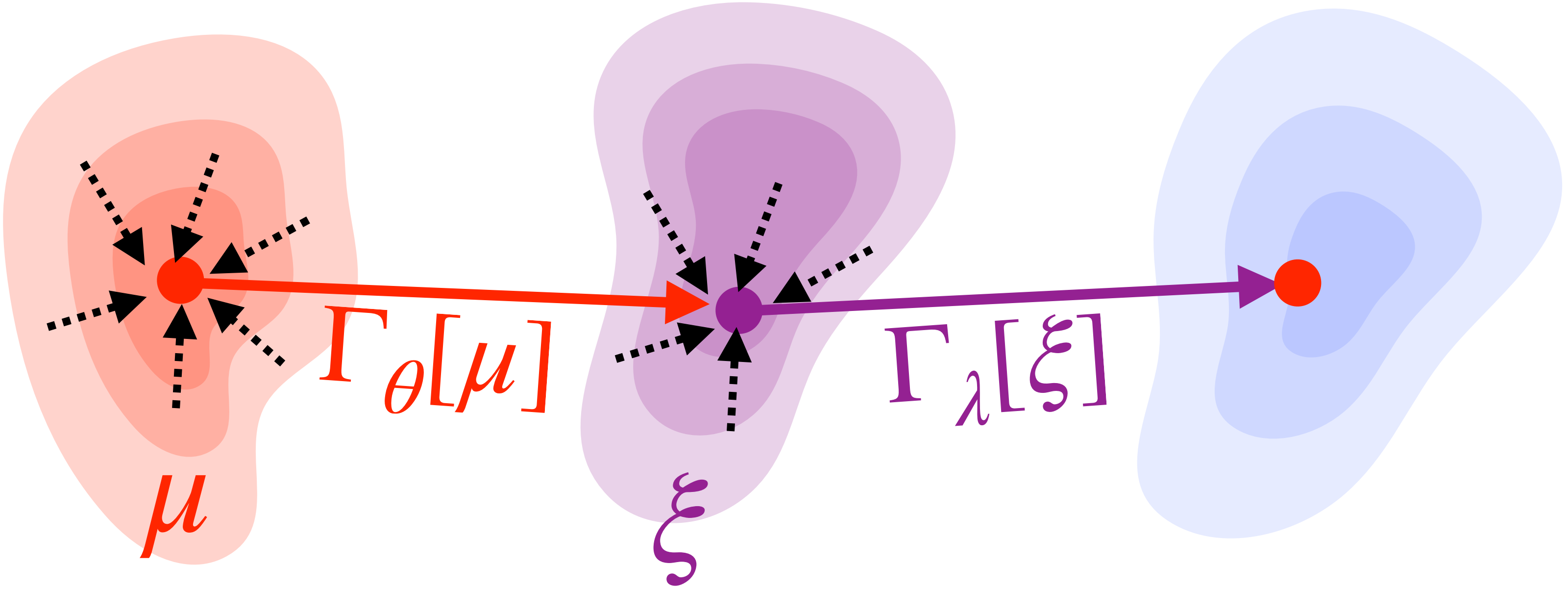
$$\mu \mapsto \Gamma_\theta[\mu]_\sharp \mu$$

Attentions Operating over Measures

Number n of token is arbitrary. (Unmasked) attention is permutation invariant.

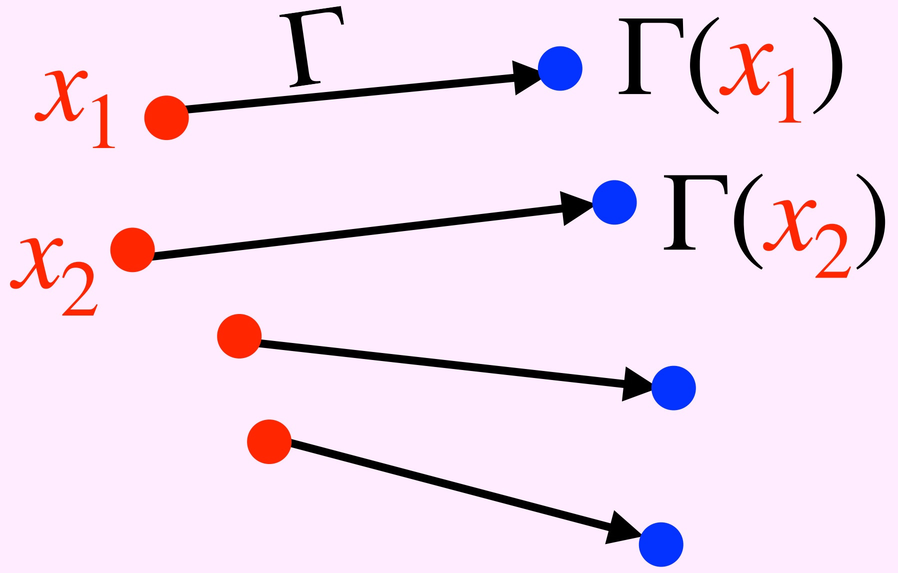
$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j\rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell\rangle}} Vx_j \xrightarrow{\boxed{\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}}} \Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y)} Vy \, \mathrm{d}\mu(y)$$



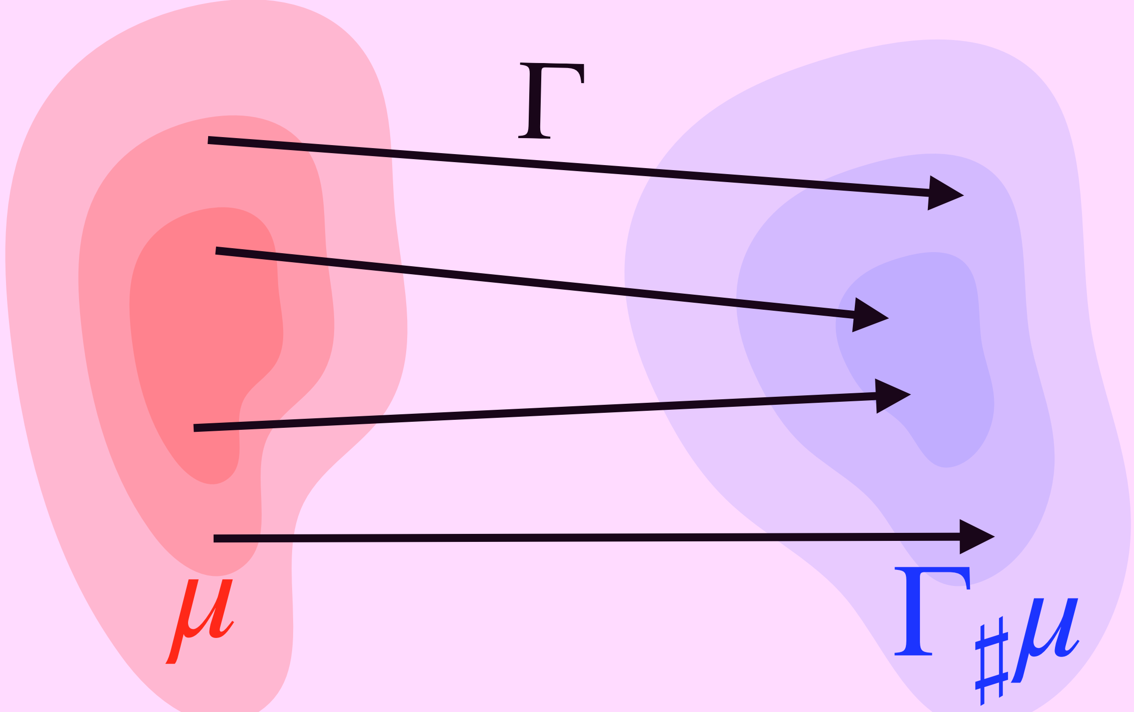


Push-forward

$$\Gamma_\sharp \sum_i \delta_{x_i} := \sum_i \delta_{\Gamma(x_i)}$$



$$(\Gamma_\sharp \mu)(B) := \mu(\Gamma^{-1}(B))$$



Attention layers

$$X \mapsto \left\{\Gamma_\theta[X](x_i)\right\}_{i=1}^n$$

$$\mu \mapsto \Gamma_\theta[\mu]_\sharp \mu$$

Composing layers

$$(\Gamma_\lambda \diamond \Gamma_\theta)[X] := \Gamma_\lambda[Y] \circ \Gamma_\theta[X]$$

where $Y := (\Gamma_\theta[X](x_i))_i$

$$(\Gamma_\lambda \diamond \Gamma_\theta)[\mu] := \Gamma_\lambda[\xi] \circ \Gamma_\theta)[\mu]$$

where $\xi := \Gamma_\theta[\mu]_\sharp \mu$

Masked Causal Attention over Measures

For NLP: architectures must be **causal** for next token prediction & generative modeling.

Masked attention mapping: $$\Gamma_\theta[X](x_i) := \sum_{j \leq i} \frac{e^{\langle Q x_i, K x_j \rangle}}{\sum_{\ell \leq i} e^{\langle Q x_i, K x_\ell \rangle}} V x_j$$

→ breaks permutation invariance.

Masked Causal Attention over Measures

For NLP: architectures must be **causal** for next token prediction & generative modeling.

Masked attention mapping: $$\Gamma_\theta[X](x_i) := \sum_{j \leq i} \frac{e^{\langle Qx_i, Kx_j \rangle}}{\sum_{\ell \leq i} e^{\langle Qx_i, Kx_\ell \rangle}} Vx_j$$

$\rightarrow$ breaks permutation invariance.

Training: next token prediction *(simplified...)* $$\min_\theta \sum_X \sum_{i=1}^{n-1} \ell(\Gamma_\theta[X](x_i), x_{i+1})$$

Testing: generative model *(simplified...)* $$X \mapsto (x_1, \ldots, x_i, \Gamma[X](x_i))$$

Masked Causal Attention over Measures

For NLP: architectures must be **causal** for next token prediction & generative modeling.

Masked attention mapping: $\Gamma_\theta[X](x_i) := \sum_{j\leq i} \frac{e^{\langle Qx_i, Kx_j\rangle}}{\sum_{\ell\leq i} e^{\langle Qx_i, Kx_\ell\rangle}} Vx_j$

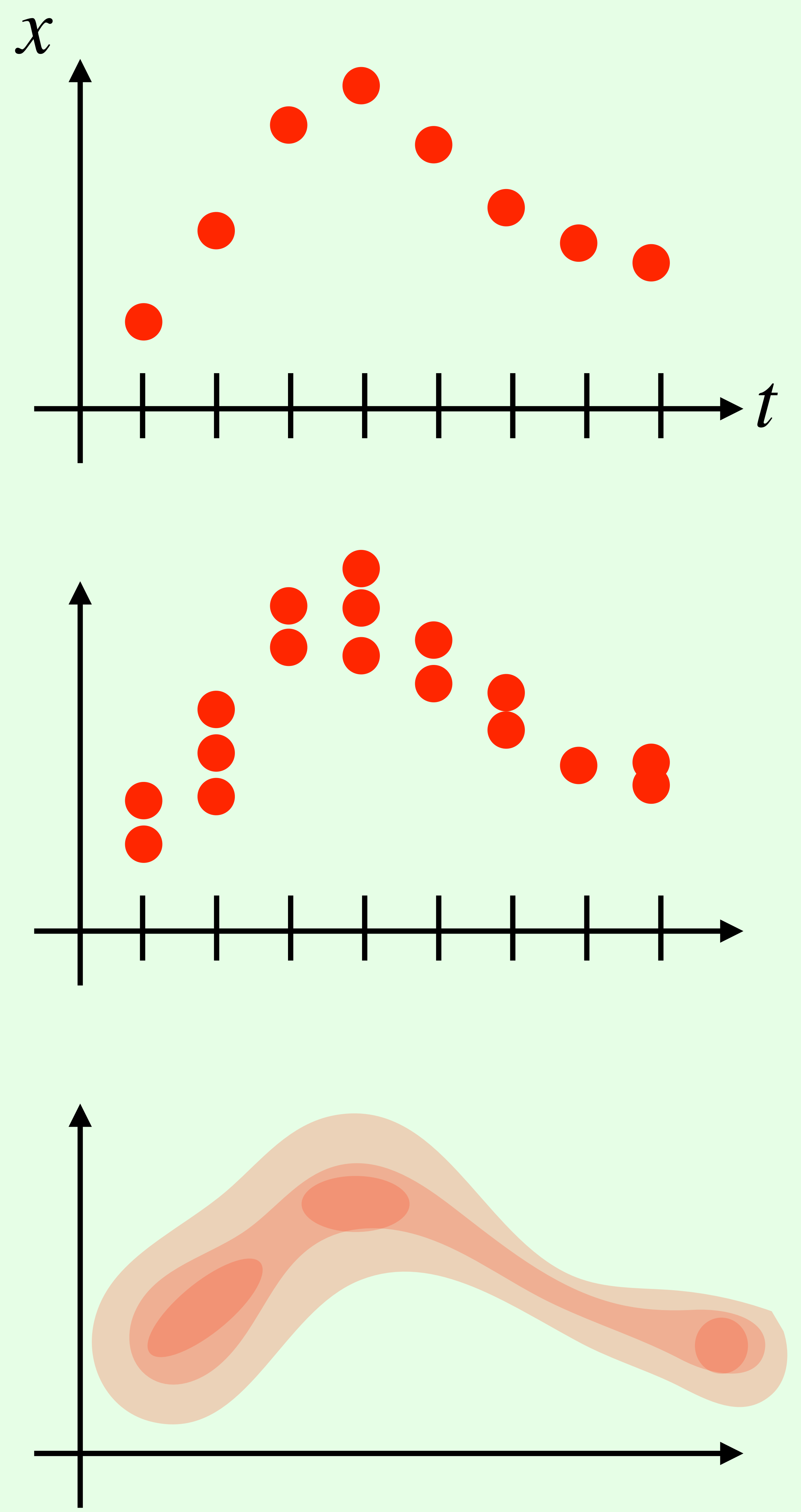
$\rightarrow$ breaks permutation invariance.

Training: next token prediction *(simplified...)* $\min_\theta \sum_X \sum_{i=1}^{n-1} \ell(\Gamma_\theta[X](x_i), x_{i+1})$

Testing: generative model *(simplified...)* $X \mapsto (x_1, \ldots, x_i, \Gamma[X](x_i))$

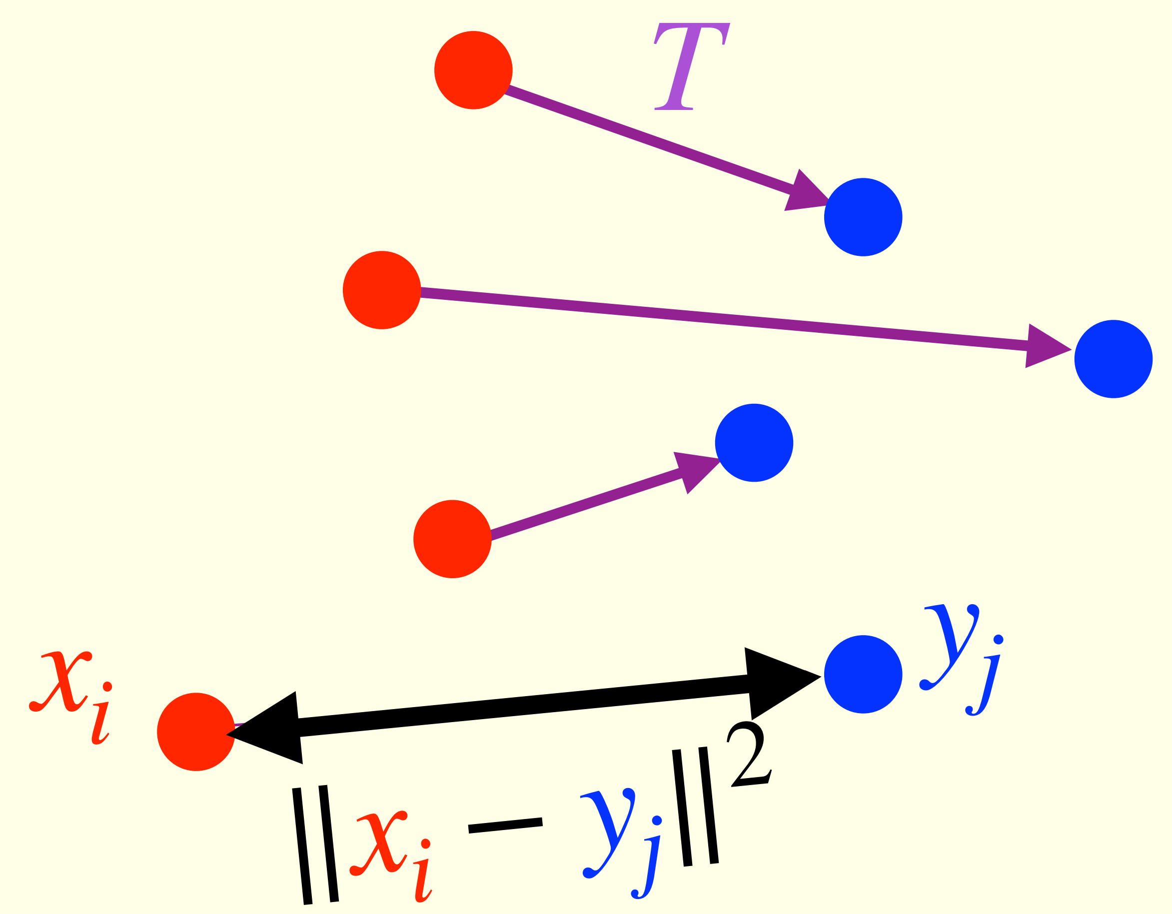
Space-time lifting: $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,t_i)}$

$$\Gamma_\theta[\mu](x,t) := \int \frac{1_{s\leq t}\, e^{\langle Qx, Ky\rangle}}{\int 1_{s'\leq t}\, e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y',s')} Vy \,\mathrm{d}\mu(y,s)$$




Arbitrary number of layers
In Context Mappings over Measures
Arbitrary number of layers
Smoothness and PDE's
Expressivity
Universality

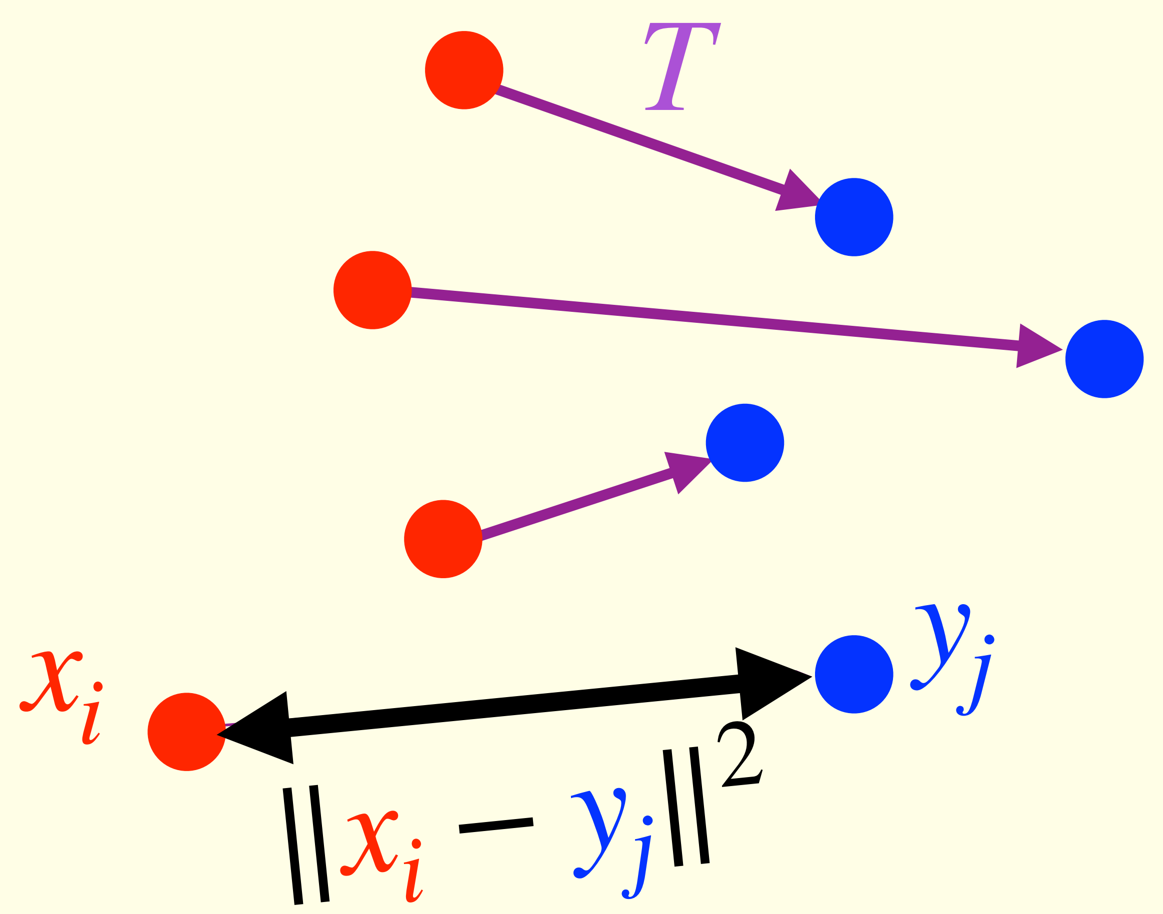
Optimal Transport (Wasserstein) Distance

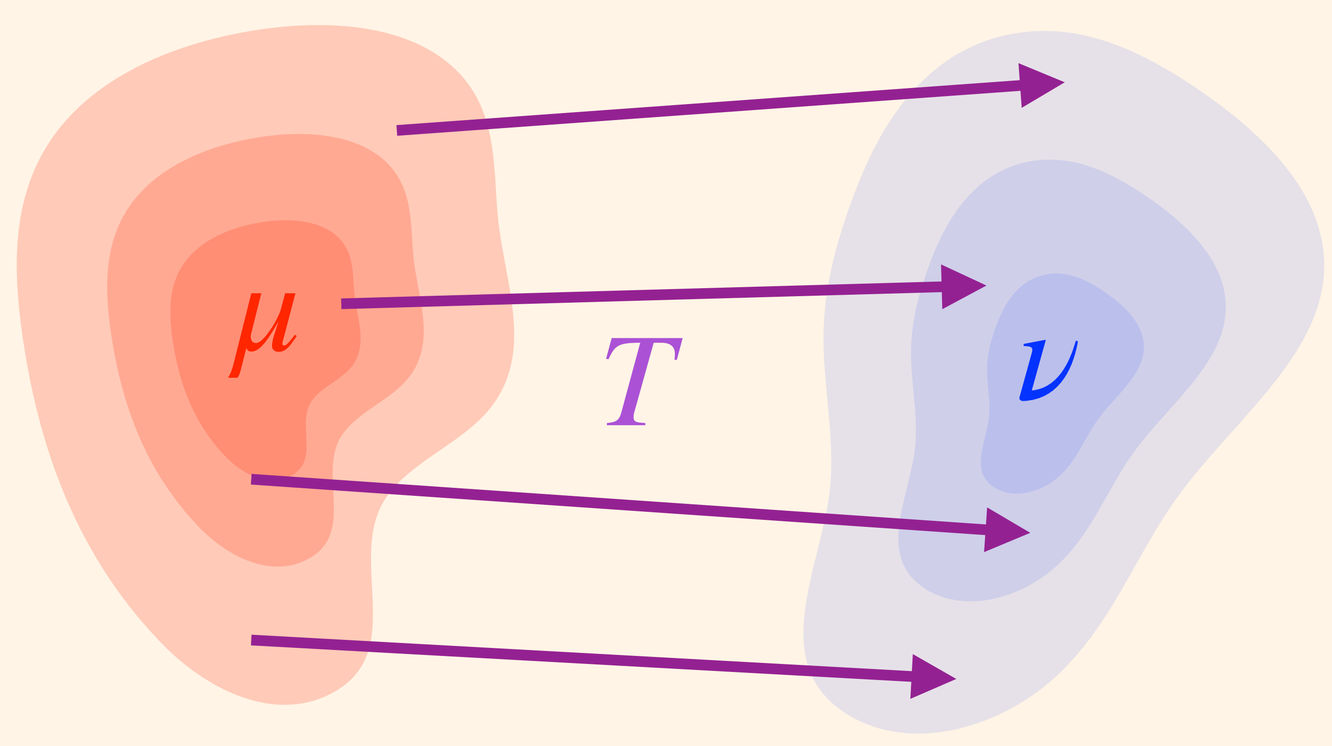


$$W_2(\mu, \nu)^2 := \min_{T} \sum_{i=1}^{n} \|x_i - y_{T(i)}\|^2$$

Monge 1784

Optimal Transport (Wasserstein) Distance

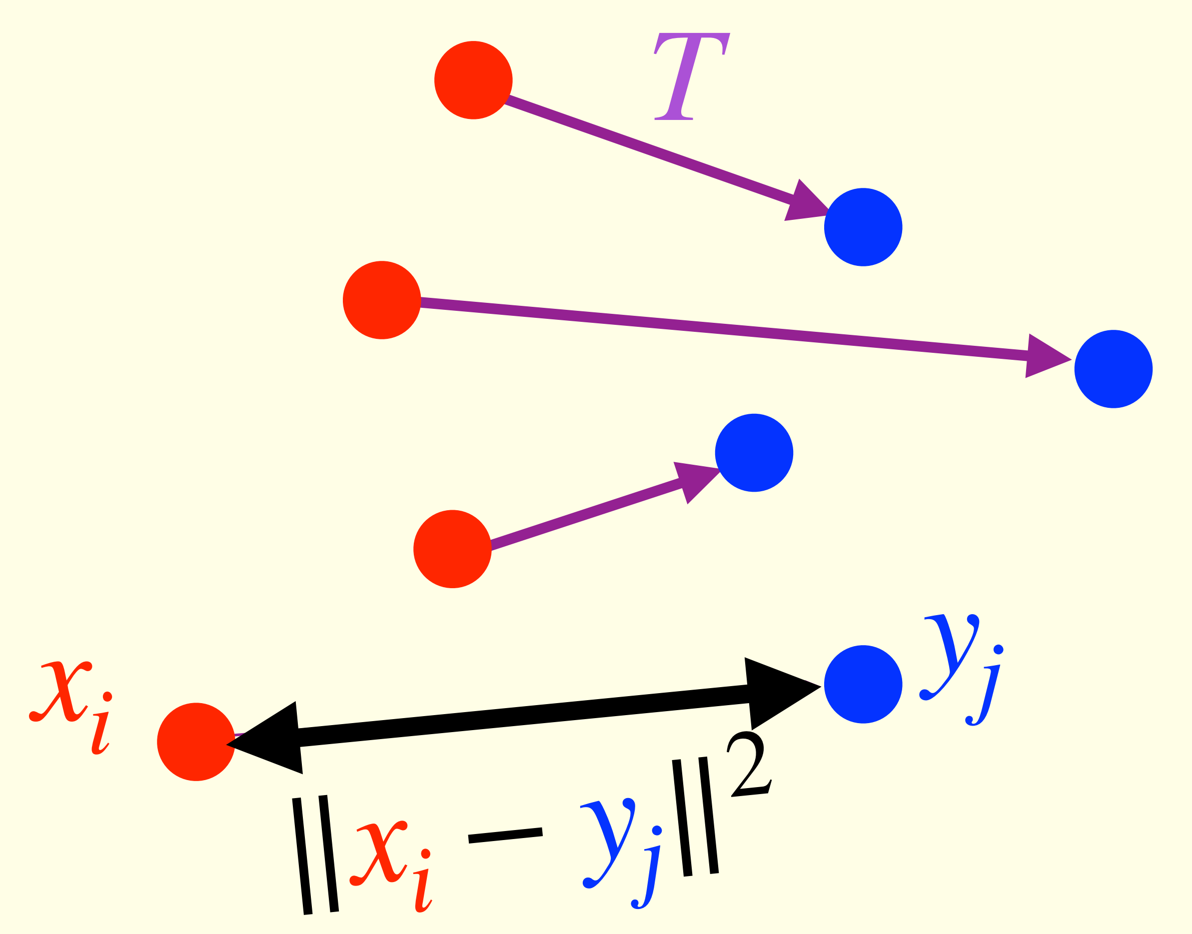


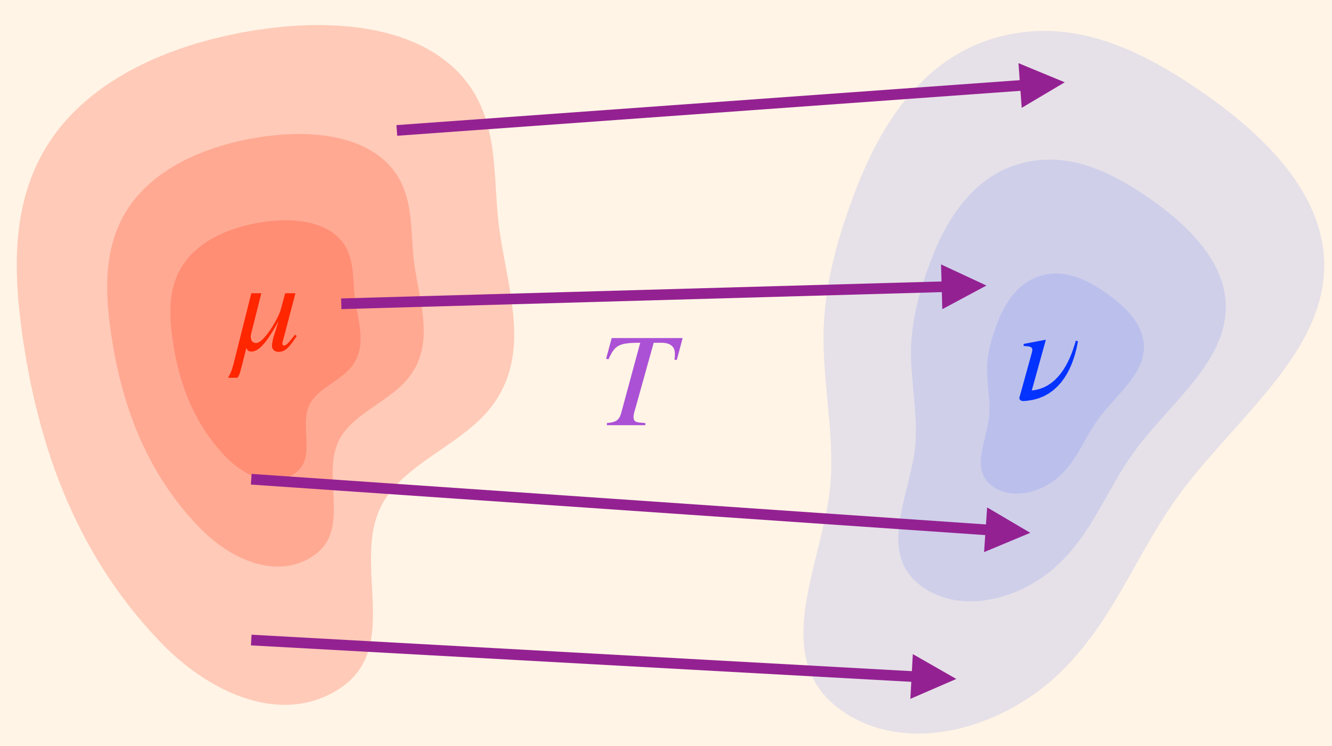


Monge 1784

$$W_2(\mu,\nu)^2 := \min_{T} \sum_{i=1}^{n} \|x_i - y_{T(i)}\|^2 \quad = \inf_{T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$

Optimal Transport (Wasserstein) Distance





Monge 1784

$$W_2(\mu, \nu)^2 := \min_{T} \sum_{i=1}^{n} \|x_i - y_{T(i)}\|^2 \quad = \inf_{T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$

Kantorovich 1942

General measures:

Kantorovich relaxation

or

Approximation by discrete measures

How Smooth is Attention?

Attention layer: $\mu \mapsto \Gamma_\theta[\mu]_\sharp \mu$

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky \rangle}}{\int e^{\langle Qx, Ky' \rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y)$$

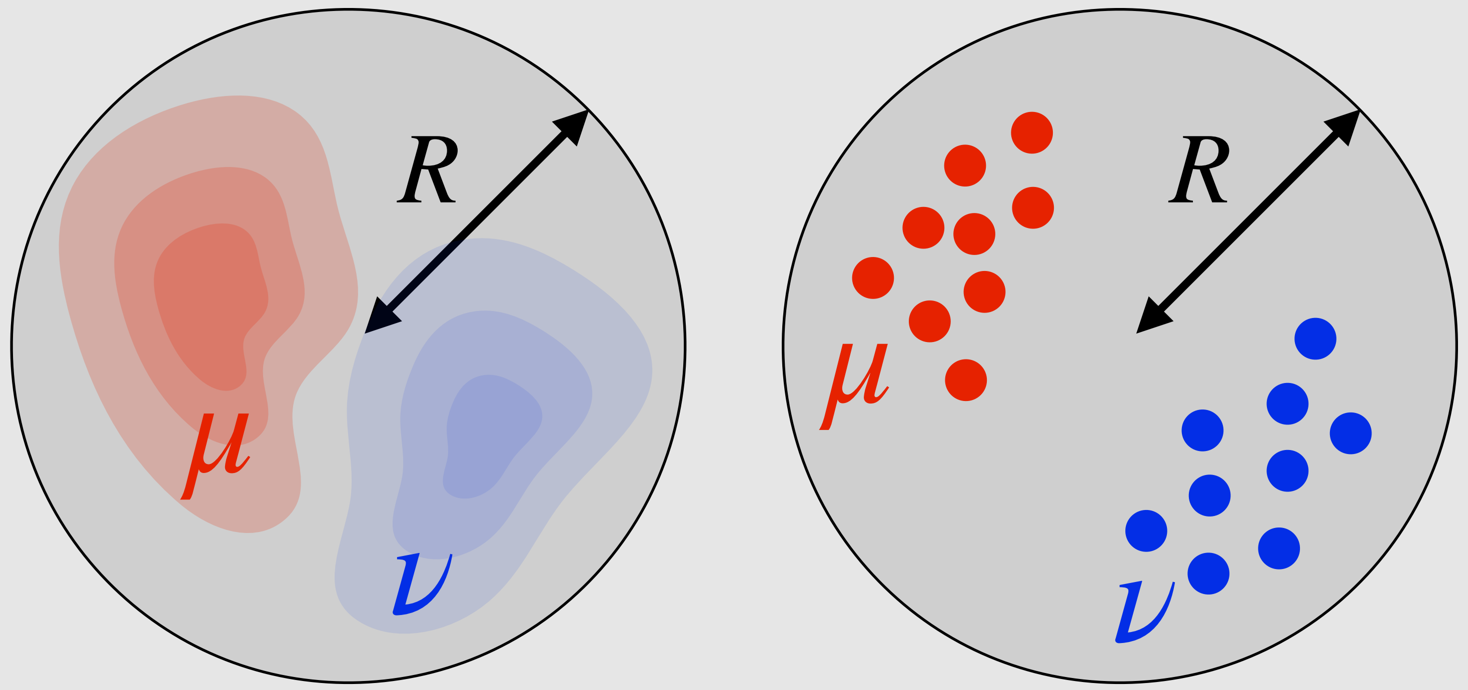
Lipschitz regularity: $\mathrm{W}_2(\Gamma_\theta[\mu]_\sharp \mu, \Gamma_\theta[\nu]_\sharp \nu) \leq C_\theta \mathrm{W}_2(\mu, \nu)$

Applications:
- Understanding robustness to attacks.
- Well-poseness of very deep transformers.

How Smooth is Attention?

Attention layer: $\mu \mapsto \Gamma_\theta[\mu]_\sharp \mu$

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky \rangle}}{\int e^{\langle Qx, Ky' \rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y)$$



Lipschitz regularity: $\mathrm{W}_2(\Gamma_\theta[\mu]_\sharp \mu, \Gamma_\theta[\nu]_\sharp \nu) \leq C_\theta \mathrm{W}_2(\mu, \nu)$

Applications:
- Understanding robustness to attacks.
- Well-poseness of very deep transformers.

Theorem: [Castin, Peyré, Ablin]

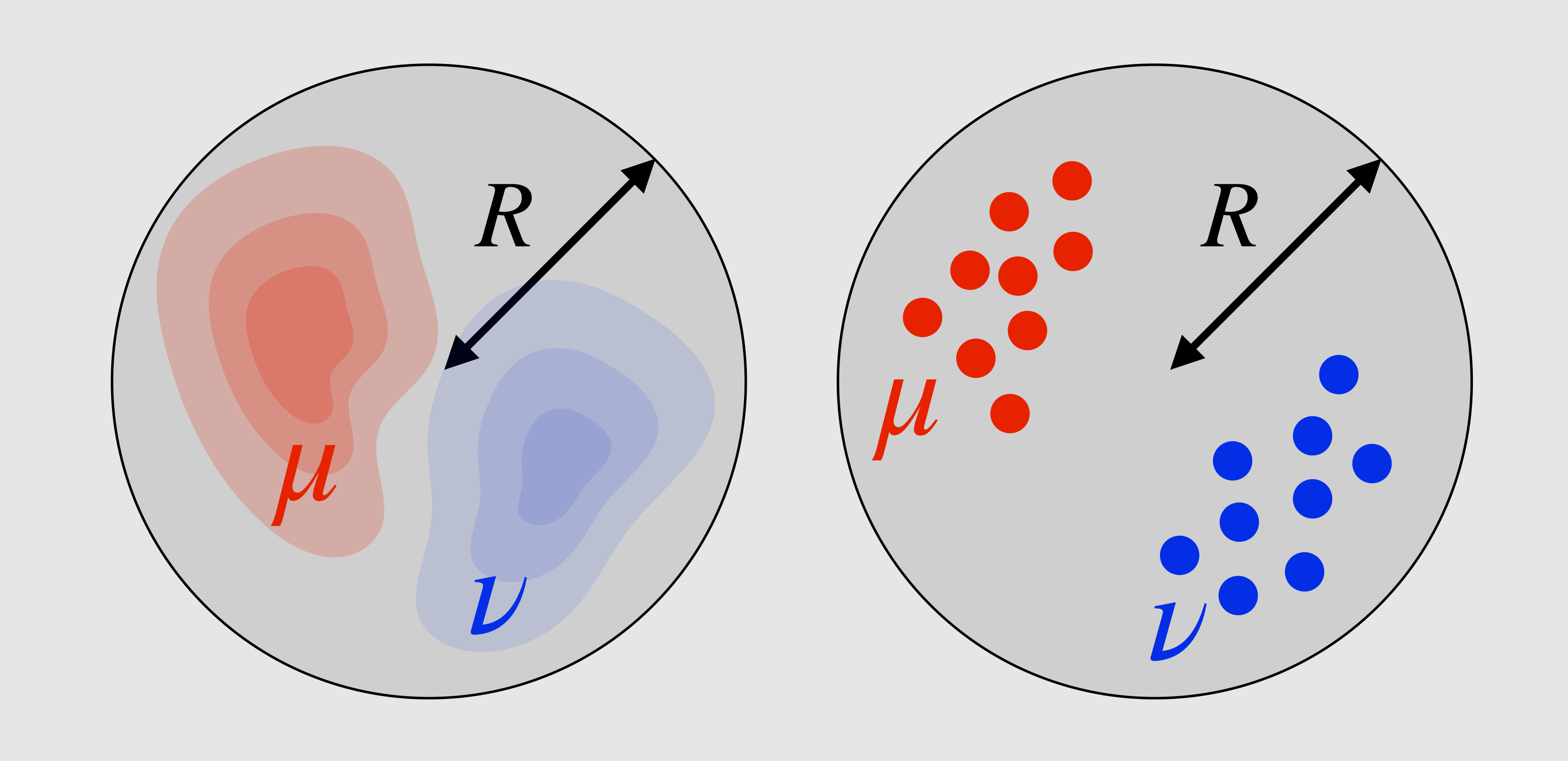
If $\mathrm{supp}(\mu), \mathrm{supp}(\nu) \subset B(0,R)$, $\quad C_\theta \leq \|V\|(1 + 3\|K^\top Q\|R^2)e^{2\|K^\top Q\|R^2}$

If furthermore $\mu = \frac{1}{n}\sum_i \delta_{x_i}, \nu = \frac{1}{n}\sum_i \delta_{y_i}$ $\quad C_\theta \leq \|V\|\|K^\top Q\|R^2\sqrt{12n + 3}$

How Smooth is Attention?

Attention layer: $\mu \mapsto \Gamma_\theta[\mu]_\sharp \mu$

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky \rangle}}{\int e^{\langle Qx, Ky' \rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y)$$



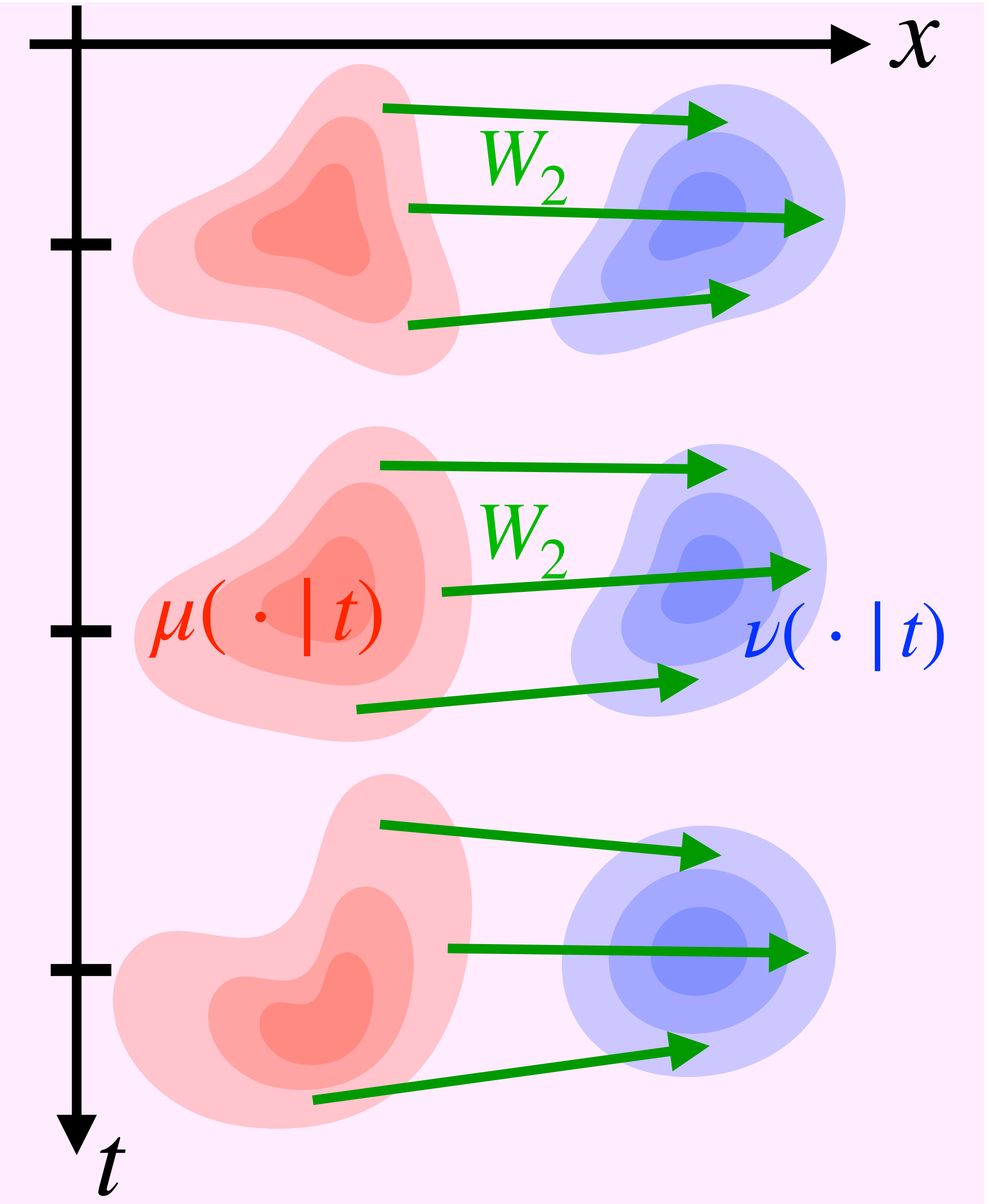
Lipschitz regularity: $\mathrm{W}_2(\Gamma_\theta[\mu]_\sharp \mu, \Gamma_\theta[\nu]_\sharp \nu) \leq C_\theta \mathrm{W}_2(\mu, \nu)$

Applications:
- Understanding robustness to attacks.
- Well-poseness of very deep transformers.

Theorem: [Castin, Peyré, Ablin]

If $\mathrm{supp}(\mu), \mathrm{supp}(\nu) \subset B(0,R)$, $\qquad C_\theta \leq \|V\|(1 + 3\|K^\top Q\|R^2) e^{2\|K^\top Q\|R^2}$

If furthermore $\mu = \frac{1}{n}\sum_i \delta_{x_i}, \nu = \frac{1}{n}\sum_i \delta_{y_i}$ $\qquad C_\theta \leq \|V\| \|K^\top Q\| R^2 \sqrt{12n + 3}$

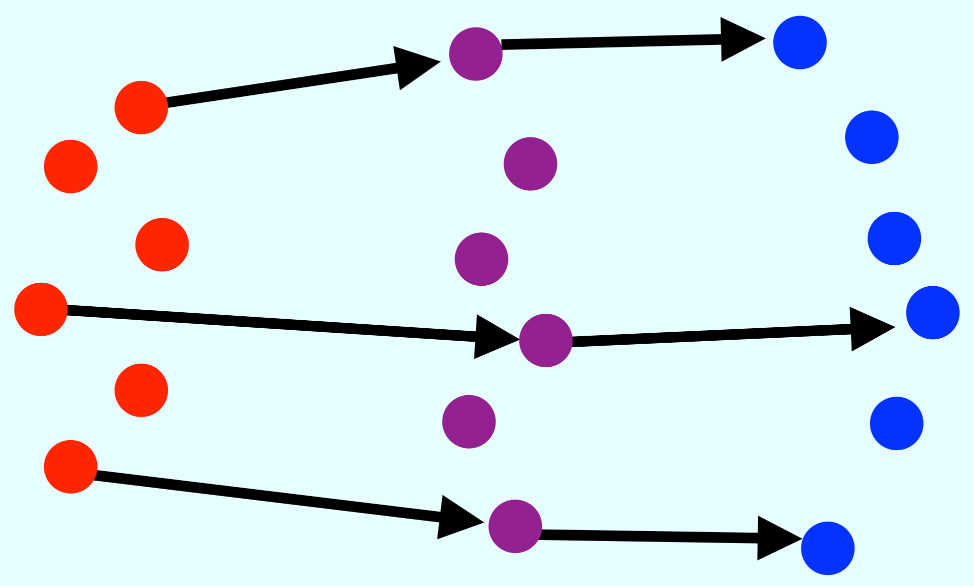


Extension to masked attention: use $W_2^{\mathrm{cond}}(\mu, \nu)^2 := \int_0^1 W_2^2(\mu(\,\cdot\,|\,t), \nu(\,\cdot\,|\,t)) \mathrm{d}\mu_{[0,1]}(t)$

Infinite Depth as a Neural PDE

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky \rangle}}{\int e^{\langle Qx, Ky' \rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta = (Q, K, V) \qquad \mu(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)}$$

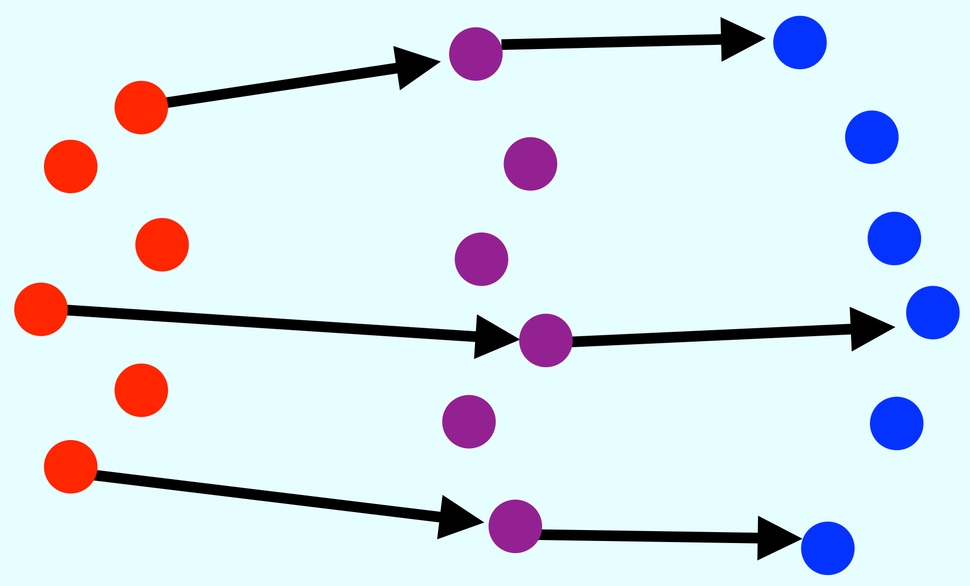
$$x_i(t+1) = x_i(t) + \frac{1}{T} \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$



Infinite Depth as a Neural PDE

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta = (Q, K, V) \qquad \mu(t) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$$

$$x_i(t+1) = x_i(t) + \frac{1}{T}\Gamma_{\theta(t)}[\mu(t)](x_i(t))$$

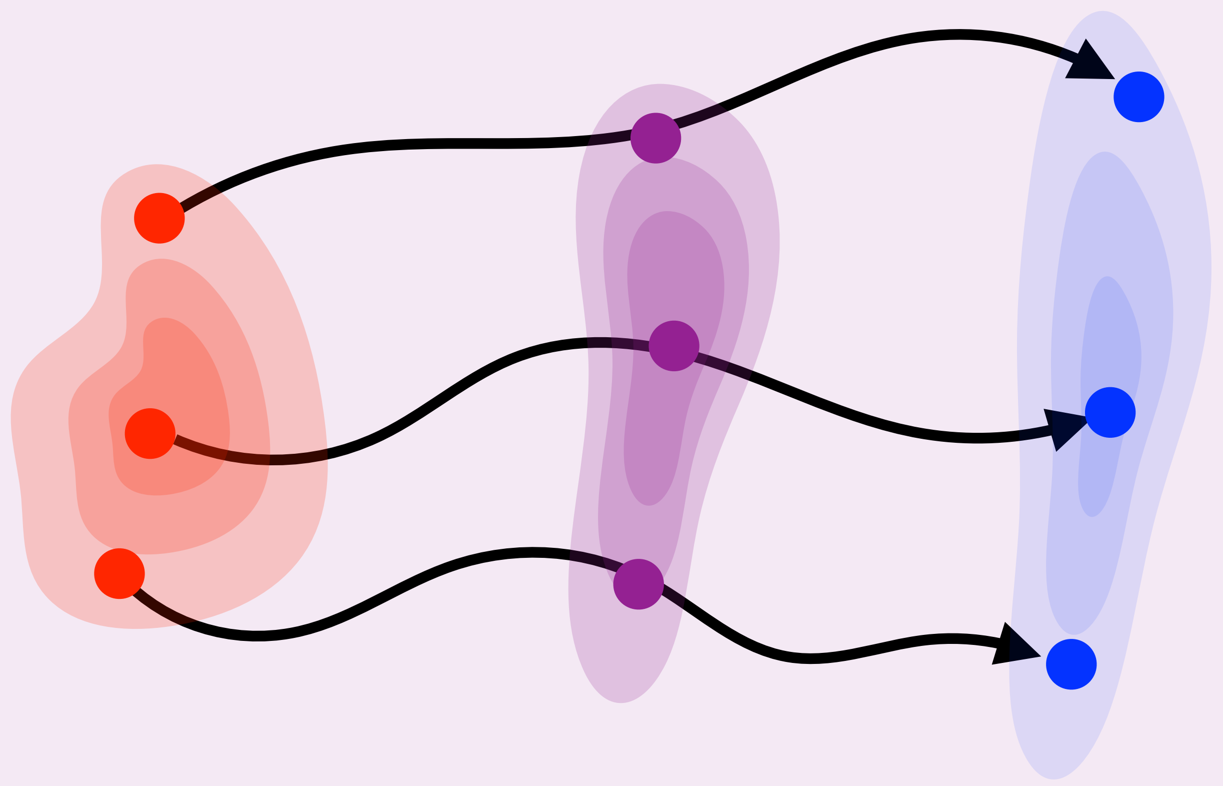


$T \to +\infty$

Infinite depth

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}(t) = \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$

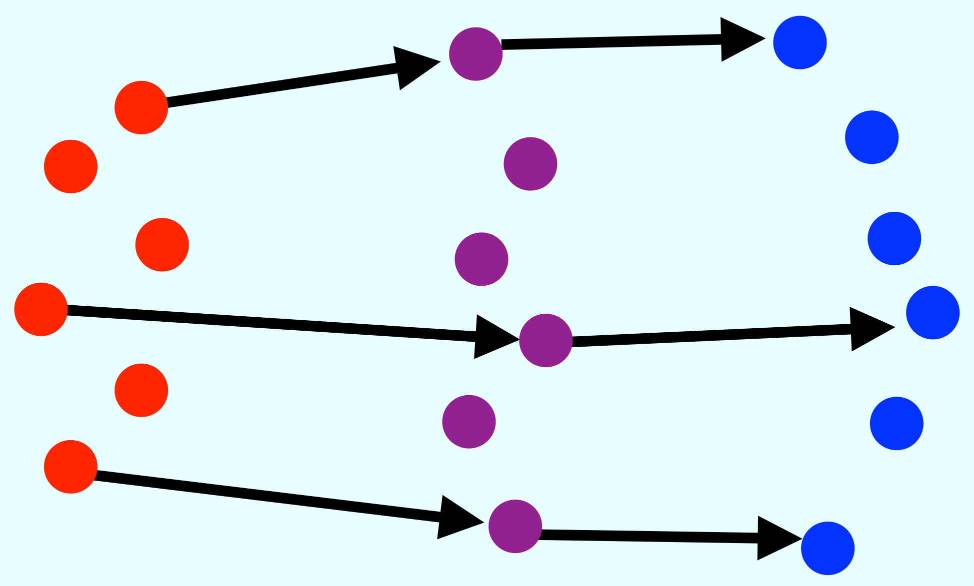
Coupled EDOs



Infinite Depth as a Neural PDE

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta = (Q, K, V) \qquad \mu(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)}$$

$$x_i(t+1) = x_i(t) + \frac{1}{T} \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$

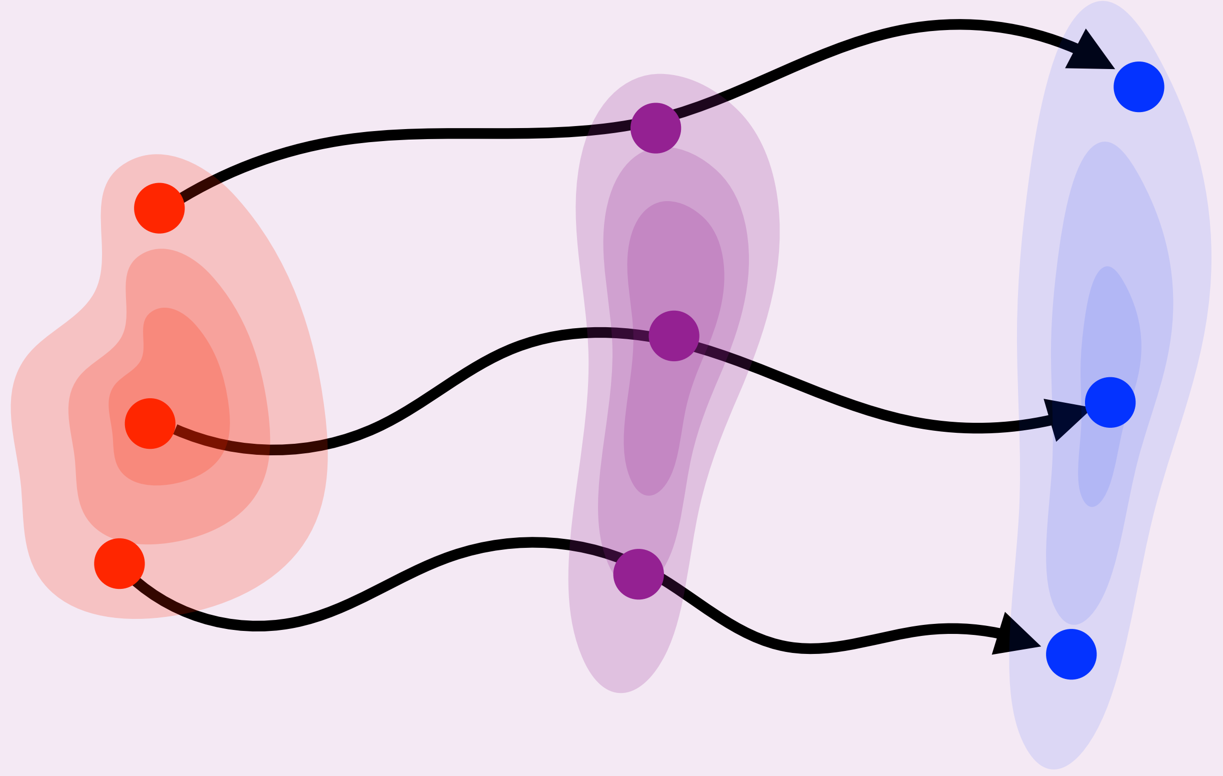


$T \to +\infty$

Infinite depth

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}(t) = \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$

Coupled EDOs



Mean field

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} + \mathrm{div}(\mu \Gamma_\theta[\mu]\,) = 0$$

Non-linear PDE

→ Not a Wasserstein flow :(

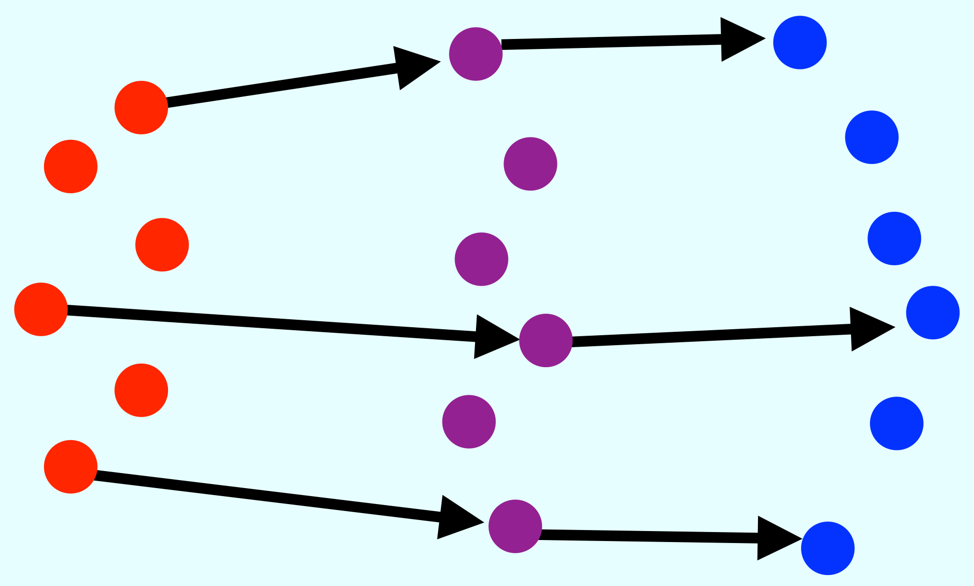
[Sander, Ablin, Blondel, Peyré, 2022]

[Geshkovski, Letrouit, Polyanskiy, Rigollet 2023]

Infinite Depth as a Neural PDE

$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta = (Q, K, V) \qquad \mu(t) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$$

$$x_i(t+1) = x_i(t) + \frac{1}{T}\Gamma_{\theta(t)}[\mu(t)](x_i(t))$$

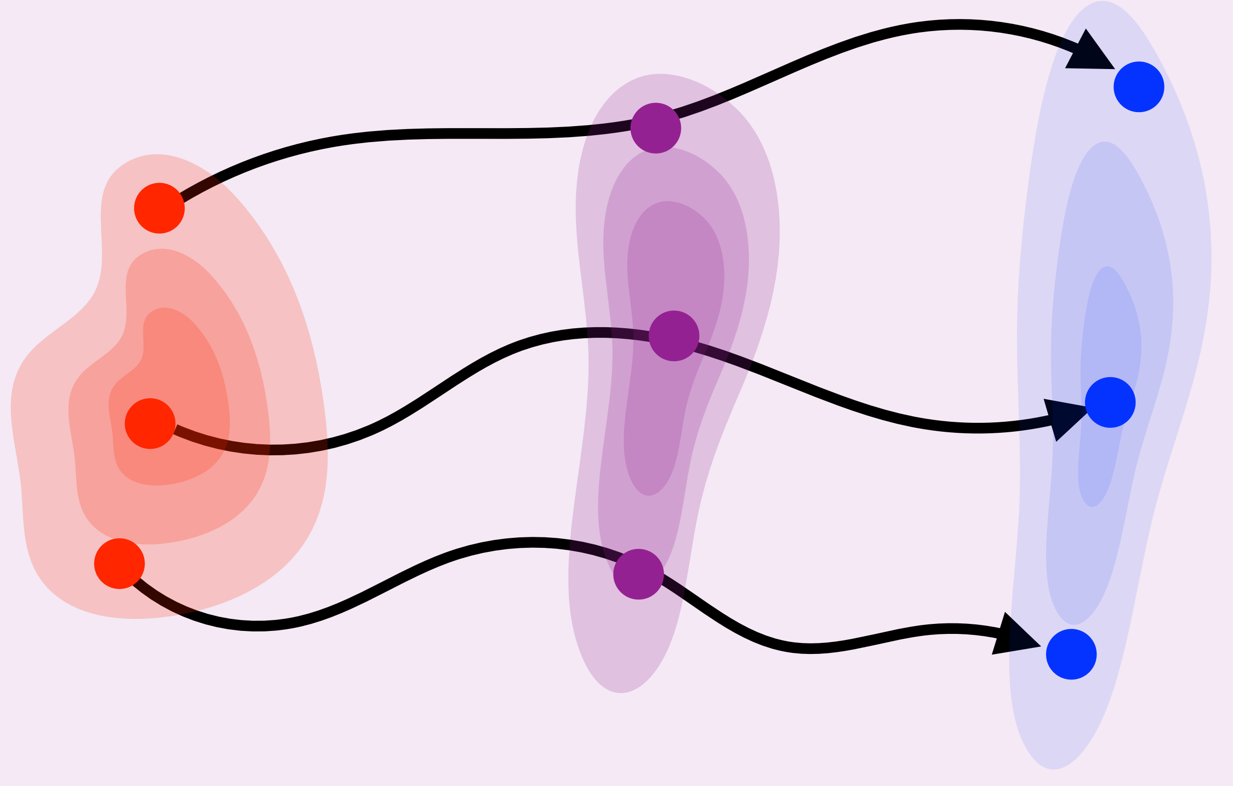


$T \to +\infty$

Infinite depth

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}(t) = \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$

Coupled EDOs



Mean field

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} + \mathrm{div}(\mu\Gamma_\theta[\mu]\,) = 0$$

Non-linear PDE

→ Not a Wasserstein flow :(

[Sander, Ablin, Blondel, Peyré, 2022]

[Geshkovski, Letrouit, Polyanskiy, Rigollet 2023]

Transformer: $T_\theta[\mu_0] : x(t=0) \xrightarrow[\mu(t=0)=\mu_0]{\dot{x} = \Gamma_\theta[\mu](x)} x(t=1)$

Training: $\min_\theta \sum_k \ell(T_\theta[\mu^k](x^k), y^k)$

Context | Previous | Next

« *Theorem* » convergence to the global minimum if
- initial loss small enough
- enough heads
- $(\mu^k)_k$ separated

→ *Talks by Pierre Marion and Raphaël Barboni*

Gaussian Case and Clustering

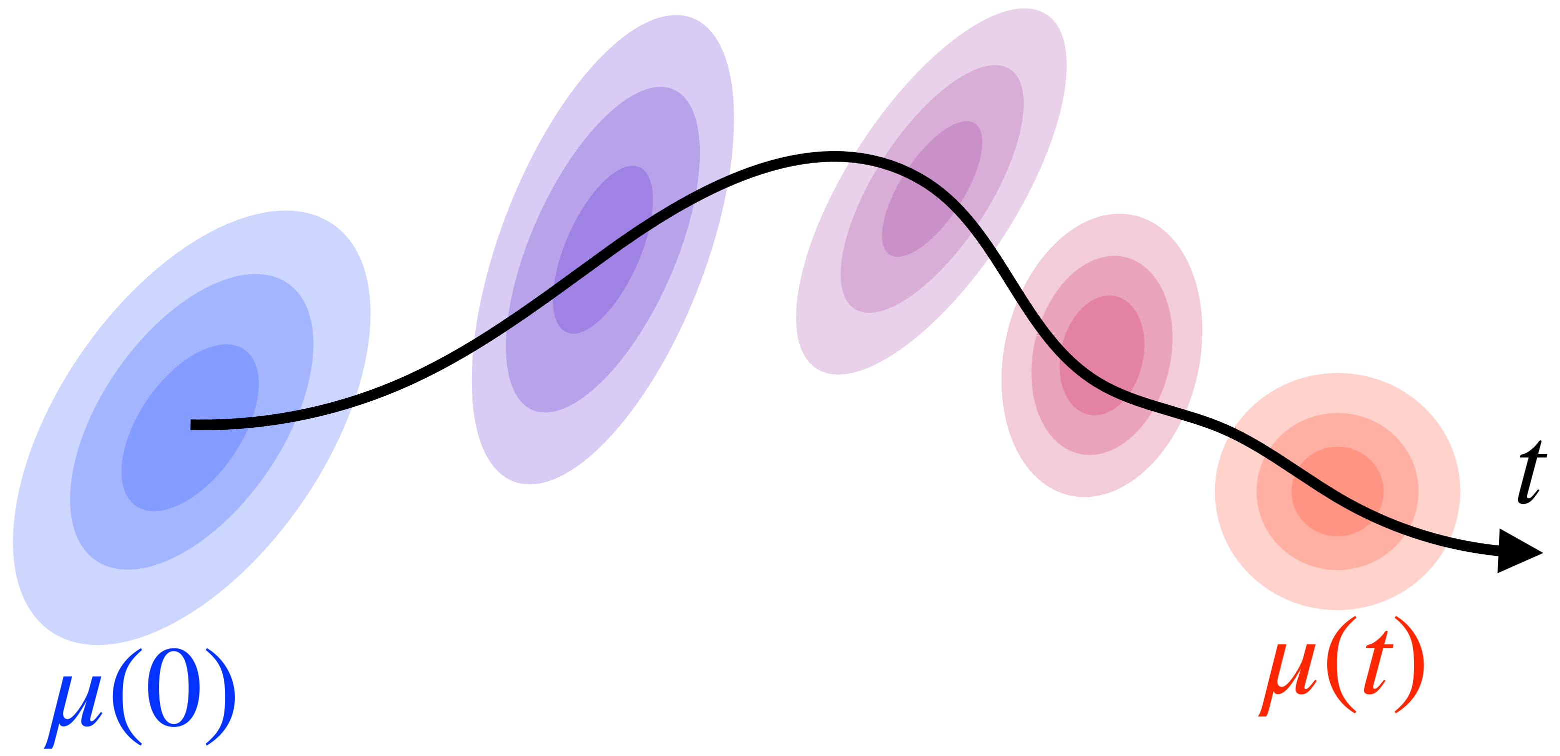
$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta(t) = (Q(t), K(t), V(t)) \qquad \frac{\mathrm{d}\mu}{\mathrm{d}t} + \mathrm{div}(\mu \Gamma_\theta[\mu]) = 0$$

Theorem [Valérie Castin]: If $\mu(0) = \mathcal{N}(m(0), \Sigma(0))$,

then $\mu(s) = \mathcal{N}(m(s), \Sigma(s))$

$$\dot{m} = V(\mathrm{Id} + \Sigma Q^\top K) m$$

$$\dot{\Sigma} = V \Sigma Q^\top K \Sigma + \Sigma K^\top Q \Sigma V^\top$$



Gaussian Case and Clustering

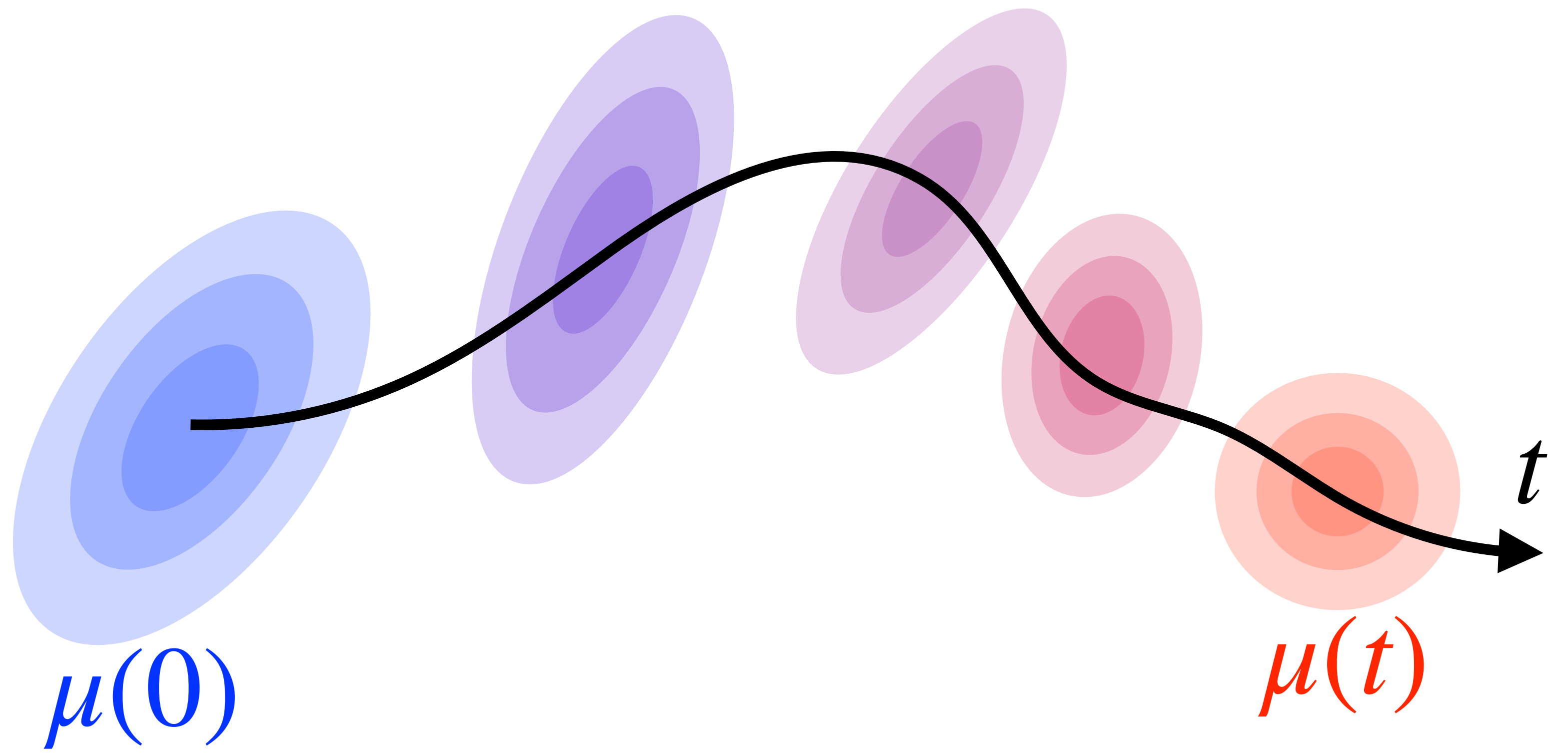
$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta(t) = (Q(t), K(t), V(t)) \qquad \frac{\mathrm{d}\mu}{\mathrm{d}t} + \mathrm{div}(\mu \Gamma_\theta[\mu]) = 0$$

Theorem [Valérie Castin]: If $\mu(0) = \mathcal{N}(m(0), \Sigma(0))$,

then $\mu(s) = \mathcal{N}(m(s), \Sigma(s))$

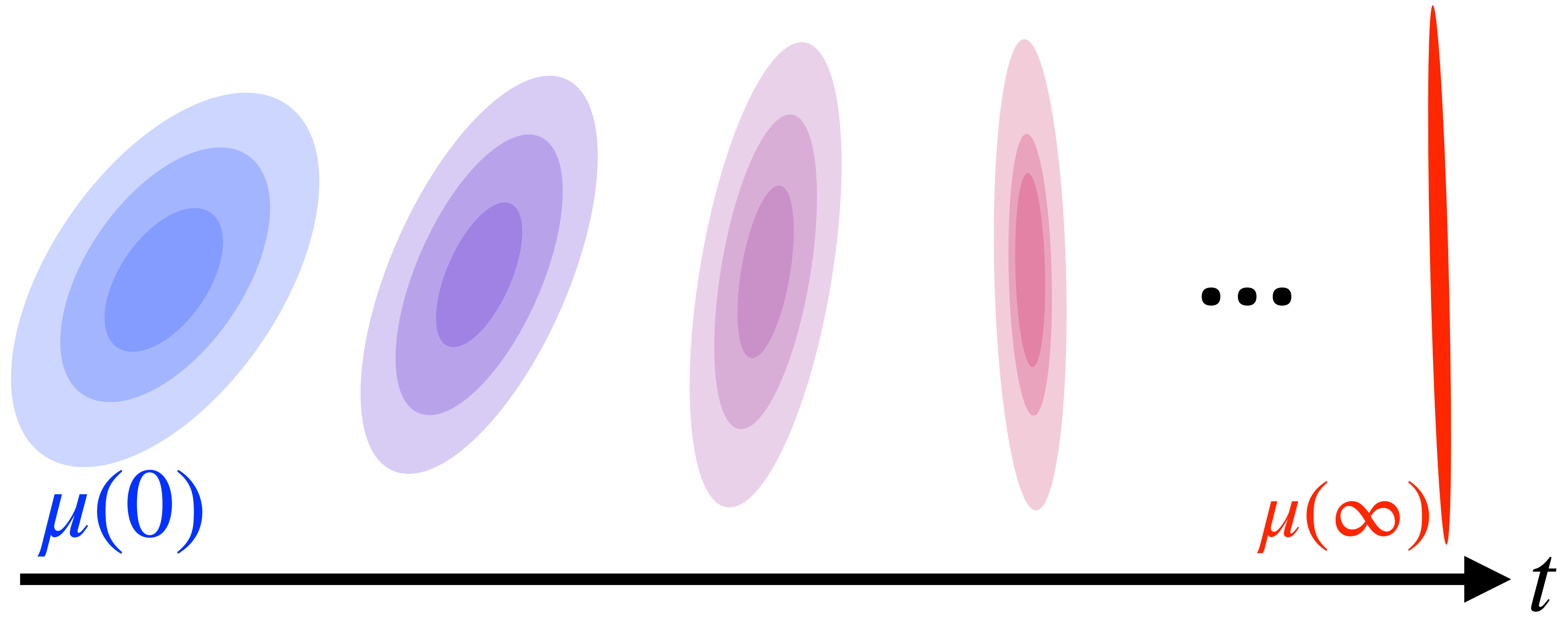
$$\dot{m} = V(\mathrm{Id} + \Sigma Q^\top K) m$$

$$\dot{\Sigma} = V \Sigma Q^\top K \Sigma + \Sigma K^\top Q \Sigma V^\top$$



Theorem [Valérie Castin]:
If $V(t) = \mathrm{Id}$ and $K(t)^\top Q(t)$ symmetric, stationary points of $\Sigma(t)$ have rank less than $d/2$.

Conjecture: low-rank stationary covariances for any K, Q, V.



[Geshkovski, Letrouit, Polyanskiy, Rigollet 2023]
→ The attention matrix converges to low-rank.
→ Clustering of μ for un-normalized attention.

Arbitrary number of layers
In Context Mappings over Measures
Arbitrary number of layers
Smoothness and PDE's
Expressivity
Universality

Universality

$$\Gamma_\theta[\mu](x) := x + \sum_{h=1}^{H} \int \frac{e^{\langle Q^h x, K^h y\rangle}}{\int e^{\langle Q^h x, K^h y'\rangle} \mathrm{d}\mu(y')} V^h y \, \mathrm{d}\mu(y) \quad \text{or} \qquad \Gamma_\theta[\mu](x) := \mathrm{MLP}_\theta(x)$$

Theorem [Furuya, de Hoop, Peyré]:

Let $\Gamma^\star : \mathscr{P}(\Omega) \times \Omega \to \mathbb{R}^d$ be $\mathrm{Wass}_2 \times \ell^2$-continuous on a compact $\Omega \subset \mathbb{R}^d$.

For any ε there exists N and $(\theta_1, \ldots, \theta_N)$ such that

$$\forall (\mu, x) \in \mathscr{P}(\Omega) \times \Omega, |\,\Gamma^\star[\mu](x) - \Gamma_{\theta_N} \diamond \cdots \diamond \Gamma_{\theta_1}[\mu](x)\,| \leq \varepsilon$$

with token dimensions $\leq 4d$ and $H \leq d$.

Novelties:
fixed dimensions,
arbitrary # tokens.

Masked transformers:
requires Lipschitz
in time.

Universality

$$\Gamma_\theta[\mu](x) := x + \sum_{h=1}^{H} \int \frac{e^{\langle Q^h x, K^h y\rangle}}{\int e^{\langle Q^h x, K^h y'\rangle} \mathrm{d}\mu(y')} V^h y \, \mathrm{d}\mu(y) \quad \text{or} \qquad \Gamma_\theta[\mu](x) := \mathrm{MLP}_\theta(x)$$

Theorem [Furuya, de Hoop, Peyré]:

Let $\Gamma^\star : \mathscr{P}(\Omega) \times \Omega \to \mathbb{R}^d$ be $\mathrm{Wass}_2 \times \ell^2$-continuous on a compact $\Omega \subset \mathbb{R}^d$.

For any ε there exists N and $(\theta_1, \ldots, \theta_N)$ such that

$$\forall (\mu, x) \in \mathscr{P}(\Omega) \times \Omega, |\, \Gamma^\star[\mu](x) - \Gamma_{\theta_N} \diamond \cdots \diamond \Gamma_{\theta_1}[\mu](x) \,| \leq \varepsilon$$

with token dimensions $\leq 4d$ and $H \leq d$.

Novelties:
fixed dimensions,
arbitrary # tokens.

Masked transformers:
requires Lipschitz
in time.

Previous works:

[Yun, Bhojanapalli, Singh Rawat, Reddi, Kumar, 2019] $\rightarrow H = 2$, dimension $\sim$ #tokens

[Agrachev, Letrouit 2019] $\rightarrow$ abstract genericity hypothesis (Lie algebra/control)

Discrete tokens: transformers are universal Turing machines: e.g. [Elhage et al 2021]

Sketch of proof

1-D elementary block: $\gamma_\theta[\mu](x) := \langle x, u\rangle + \int \frac{e^{\langle Ax+b, y\rangle}}{\int e^{\langle Ax+b,y\rangle}\mathrm{d}\mu(y)}(\langle v, y\rangle + c)\mathrm{d}\mu(y)$ $\theta := (A, b, c, u, v)$

→ First component of Attention ∘ MLP with skip connnexion.

Cylindrical algebra: $\mathscr{A} := \mathrm{Span} \bigcup_N \{\gamma_{\theta_1} \odot \cdots \odot \gamma_{\theta_N} : (\theta_1, \ldots, \theta_N)\}$ $(\gamma_1 \odot \gamma_2)[\mu](x) := \gamma_1[\mu](x)\gamma_2[\mu](x)$

Sketch of proof

1-D elementary block: $\gamma_\theta[\mu](x) := \langle x, u\rangle + \int \frac{e^{\langle Ax+b,y\rangle}}{\int e^{\langle Ax+b,y\rangle}\mathrm{d}\mu(y)}(\langle v, y\rangle + c)\mathrm{d}\mu(y)$ $\qquad \theta := (A, b, c, u, v)$

→ First component of Attention ∘ MLP with skip connnexion.

Cylindrical algebra: $\mathscr{A} := \mathrm{Span} \bigcup_N \{\gamma_{\theta_1} \odot \cdots \odot \gamma_{\theta_N} : (\theta_1, \ldots, \theta_N)\}$ $\qquad (\gamma_1 \odot \gamma_2)[\mu](x) := \gamma_1[\mu](x)\gamma_2[\mu](x)$

Proposition: any map $(\mu, x) \to (\alpha_1[\mu](x), \ldots, \alpha_d[\mu](x)) \in \mathbb{R}^d$ with $\alpha_i \in \mathscr{A}$ can be uniformly approximated by a transformer with skip connexions.

Proof sketch:

Use 1D dimension by dimension → requires $H = d$ heads.

Multiplications $\odot$ $\xrightarrow[\text{Use MLPs}]{\text{double dimension}}$ Compositions $\circ$

Compositions $\circ$ $\xrightarrow[\text{dimension.}]{\text{double}}$ In-context $\diamond$

→ Embedding dimenson = $4d$

Sketch of Proof

$$\gamma_\theta[\mu](x) := \langle x, u\rangle + \int \frac{e^{\langle Ax+b, y\rangle}}{\int e^{\langle Ax+b, y'\rangle}\mathrm{d}\mu(y')}(\langle v, y\rangle + c)\mathrm{d}\mu(y) \qquad \mathscr{A} := \mathrm{Span}\bigcup_N \{\gamma_{\theta_1} \odot \cdots \odot \gamma_{\theta_N}\}$$

Lemma: $\mathscr{A}$ is dense in continuous maps $\mathscr{P}(\Omega) \times \Omega \to \mathbb{R}$ for $\mathrm{Wass}_2 \times \ell^2$

Sketch of Proof

$$\gamma_\theta[\mu](x) := \langle x, u\rangle + \int \frac{e^{\langle Ax+b, y\rangle}}{\int e^{\langle Ax+b, y'\rangle} \mathrm{d}\mu(y')} (\langle v, y\rangle + c) \mathrm{d}\mu(y) \qquad \mathscr{A} := \mathrm{Span} \bigcup_N \{\gamma_{\theta_1} \odot \cdots \odot \gamma_{\theta_N}\}$$

Lemma: $\mathscr{A}$ is dense in continuous maps $\mathscr{P}(\Omega) \times \Omega \to \mathbb{R}$ for $\mathrm{Wass}_2 \times \ell^2$

Proof:

Stone-Weierstrass theorem

$\mathscr{P}(\Omega) \times \Omega$ is compact.

γ_θ are continuous.

$A = b = u = v = 0$, $c = 1$:

$$\gamma_\theta[\mu] = 1$$

$$\forall \theta, \gamma_\theta[\mu](x) = \gamma_\theta[\mu'](x') \overset{?}{\Longrightarrow} (\mu, x) = (\mu', x')$$

Sketch of Proof

$$\gamma_\theta[\mu](x) := \langle x, u\rangle + \int \frac{e^{\langle Ax+b, y\rangle}}{\int e^{\langle Ax+b, y'\rangle}\mathrm{d}\mu(y')}(\langle v, y\rangle + c)\mathrm{d}\mu(y) \qquad \mathscr{A} := \mathrm{Span} \bigcup_N \{\gamma_{\theta_1} \odot \cdots \odot \gamma_{\theta_N}\}$$

Lemma: $\mathscr{A}$ is dense in continuous maps $\mathscr{P}(\Omega) \times \Omega \to \mathbb{R}$ for $\mathrm{Wass}_2 \times \ell^2$

Proof:

Stone-Weierstrass theorem

$\mathscr{P}(\Omega) \times \Omega$ is compact.

γ_θ are continuous.

$A = b = u = v = 0$, $c = 1$:
$\gamma_\theta[\mu] = 1$

$$\forall \theta, \gamma_\theta[\mu](x) = \gamma_\theta[\mu'](x') \overset{?}{\Longrightarrow} (\mu, x) = (\mu', x')$$

$c = v = 0$: $\quad \langle x, u\rangle = \langle x', u\rangle$

Sketch of Proof

$$\gamma_\theta[\mu](x) := \langle x, u\rangle + \int \frac{e^{\langle Ax+b, y\rangle}}{\int e^{\langle Ax+b, y'\rangle}\mathrm{d}\mu(y')}(\langle v, y\rangle + c)\mathrm{d}\mu(y) \qquad \mathscr{A} := \mathrm{Span}\bigcup_N \{\gamma_{\theta_1} \odot \cdots \odot \gamma_{\theta_N}\}$$

Lemma: $\mathscr{A}$ is dense in continuous maps $\mathscr{P}(\Omega) \times \Omega \to \mathbb{R}$ for $\mathrm{Wass}_2 \times \ell^2$

Proof:

Stone-Weierstrass theorem

$\mathscr{P}(\Omega) \times \Omega$ is compact.

γ_θ are continuous.

$A = b = u = v = 0$, $c = 1$:
$\gamma_\theta[\mu] = 1$

$$\forall \theta, \gamma_\theta[\mu](x) = \gamma_\theta[\mu'](x') \overset{?}{\Longrightarrow} (\mu, x) = (\mu', x')$$

$c = v = 0$: $\quad \langle x, u\rangle = \langle x', u\rangle$

$A = c = u = 0$: $\quad L_1(\mu)(b) = L_1(\mu')(b)$

In 1-D: $\quad L_k(\mu)(b) := \int \frac{e^{by}y^k v}{\int e^{by'}\mathrm{d}\mu(y')}\mathrm{d}\mu(y)$

$$L_k' = L_{k+1} - L_k L_1$$

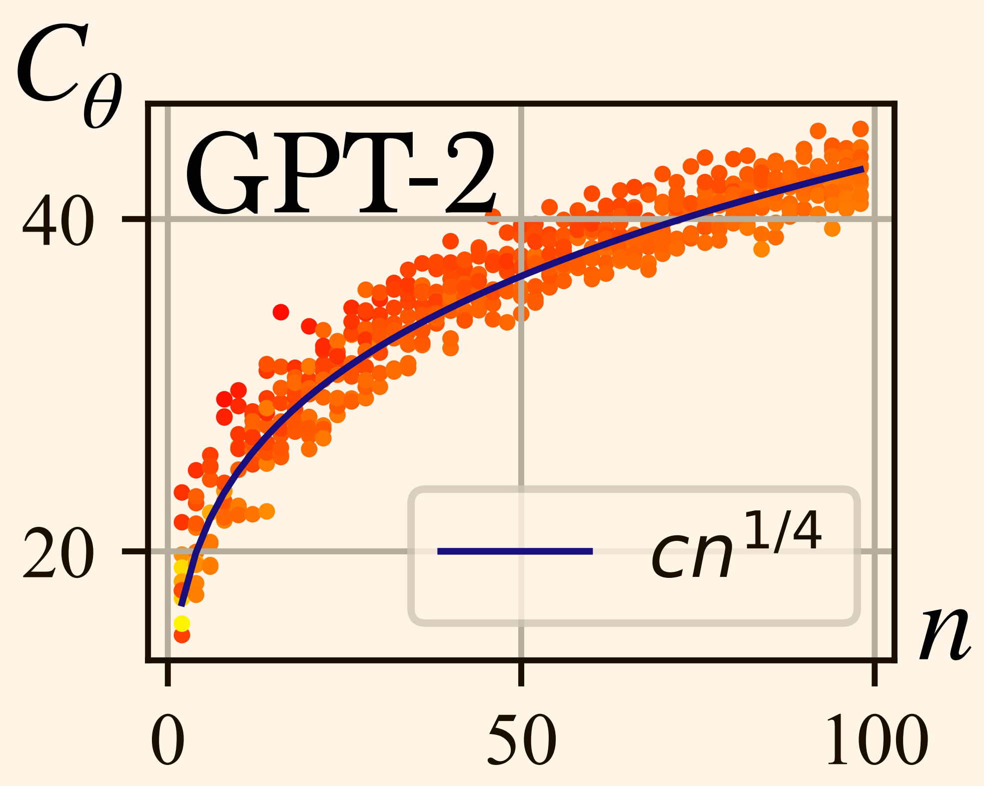
$$L_1(\mu) = L_1(\mu') \Rightarrow \forall k, L_k(\mu) = L_k(\mu') \Rightarrow \forall k, \int y^k \mathrm{d}\mu(y) = \int y^k \mathrm{d}\mu'(y)$$

In higher dimensions: use Radon transform.

Open Problems

Smoothness: bridge the gap

mean-field e^R ⟷ discrete $\sqrt{n}$ ⟷ practice $n^{1/4}$



Universality:
- Replace scalar-valued cylindrical maps by more effective functions.
- Toward quantitative approximation bound, leverage smoothness.

Optimisation:
- Understand the structure of optimal (Q, K, V)
- Why is Adam normalization needed for training?