Universal Approximation:

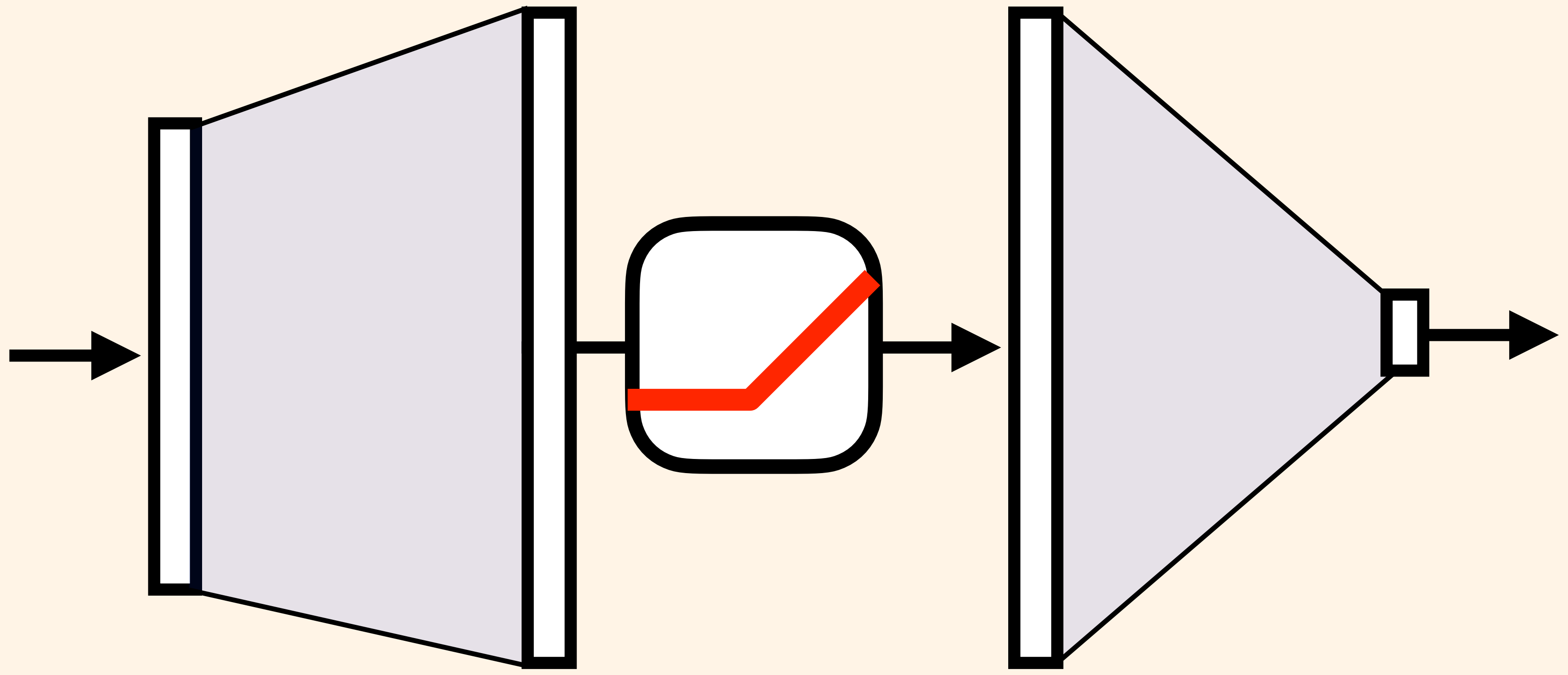
from Neural Networks to Transformers

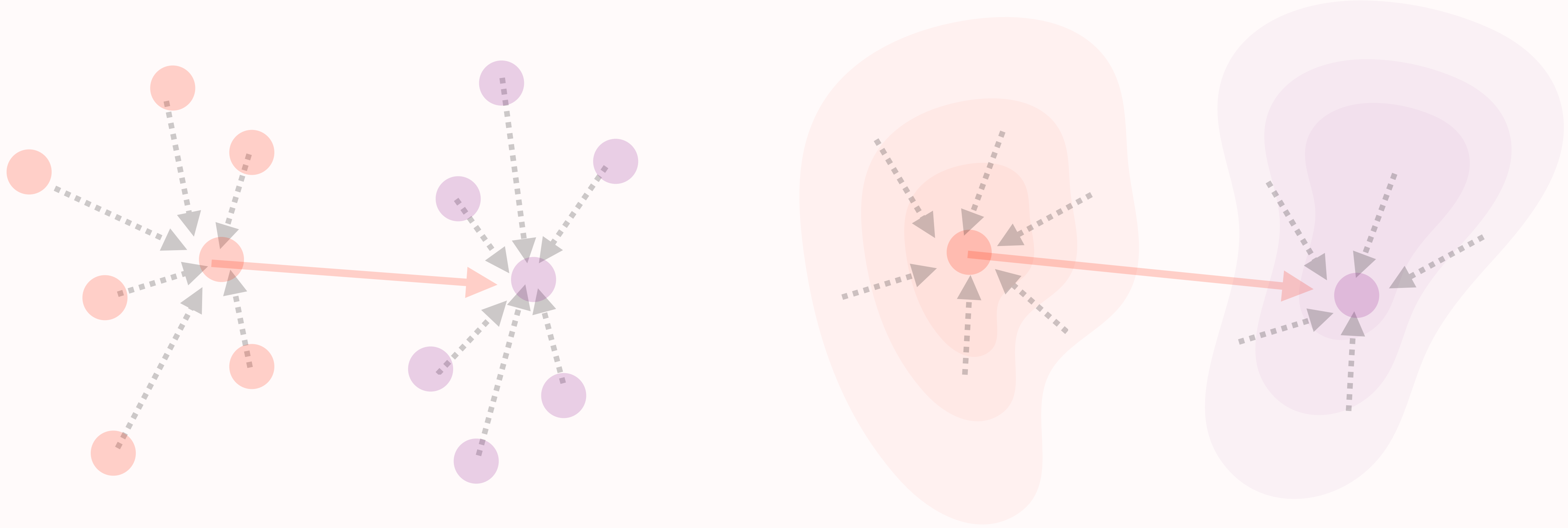
Gabriel Peyré

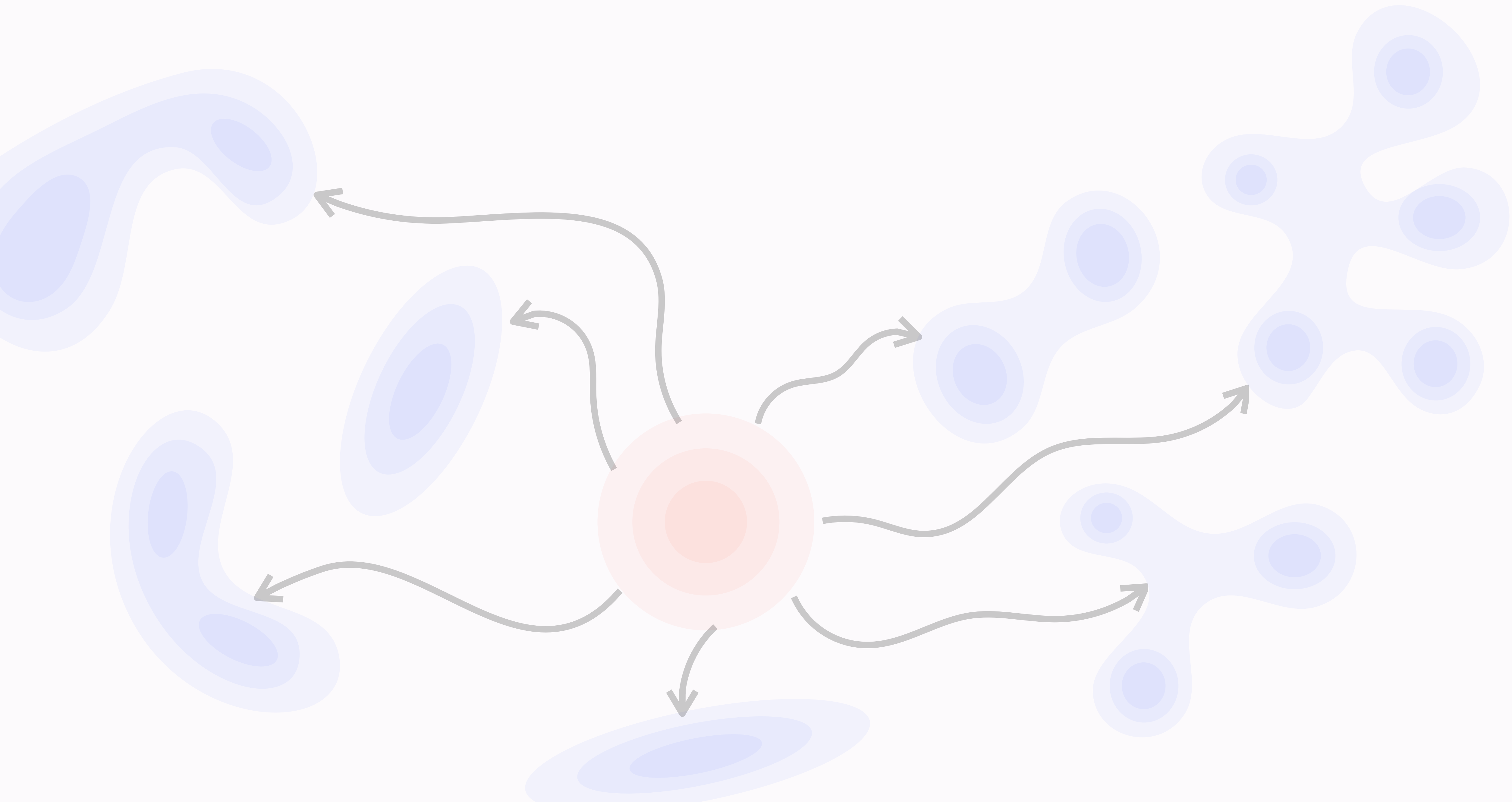
cnrs

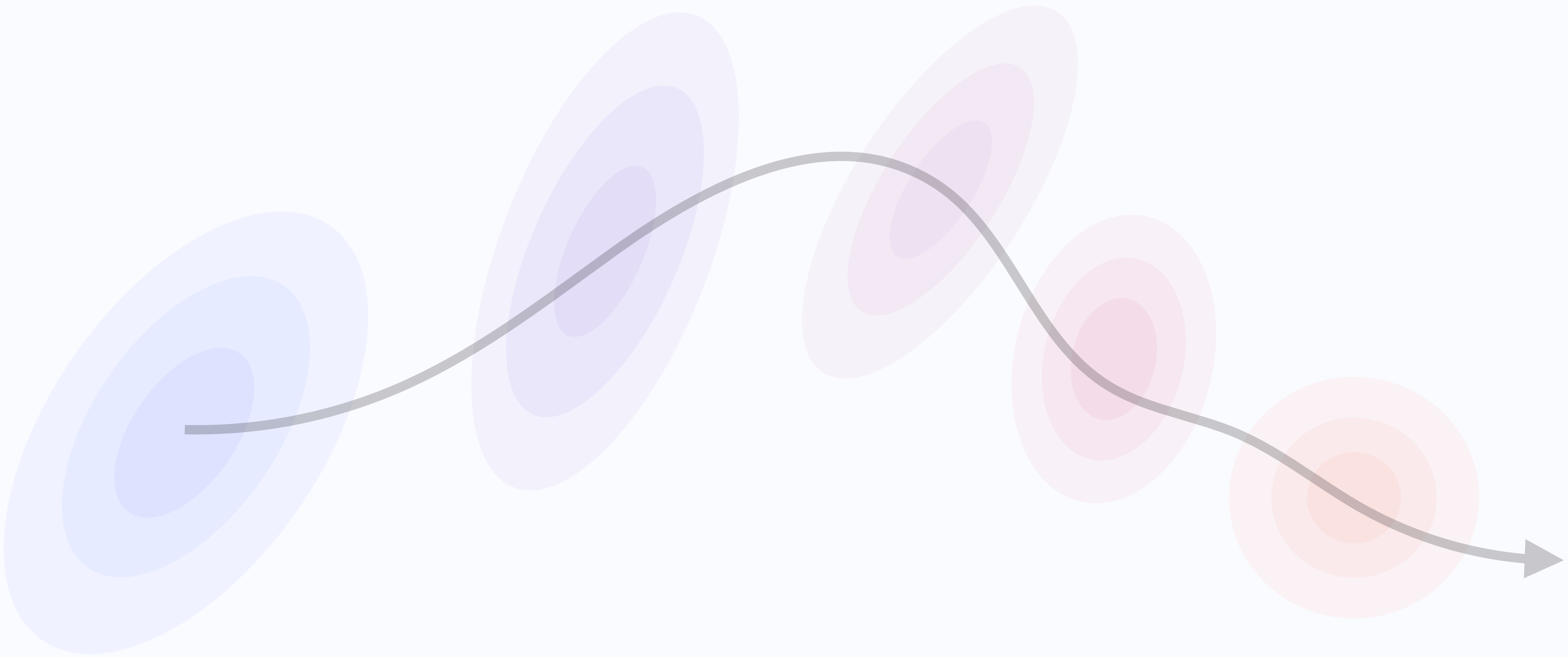
1794
ENS
ÉCOLE NORMALE SUPÉRIEURE

Perceptrons: context-free universality

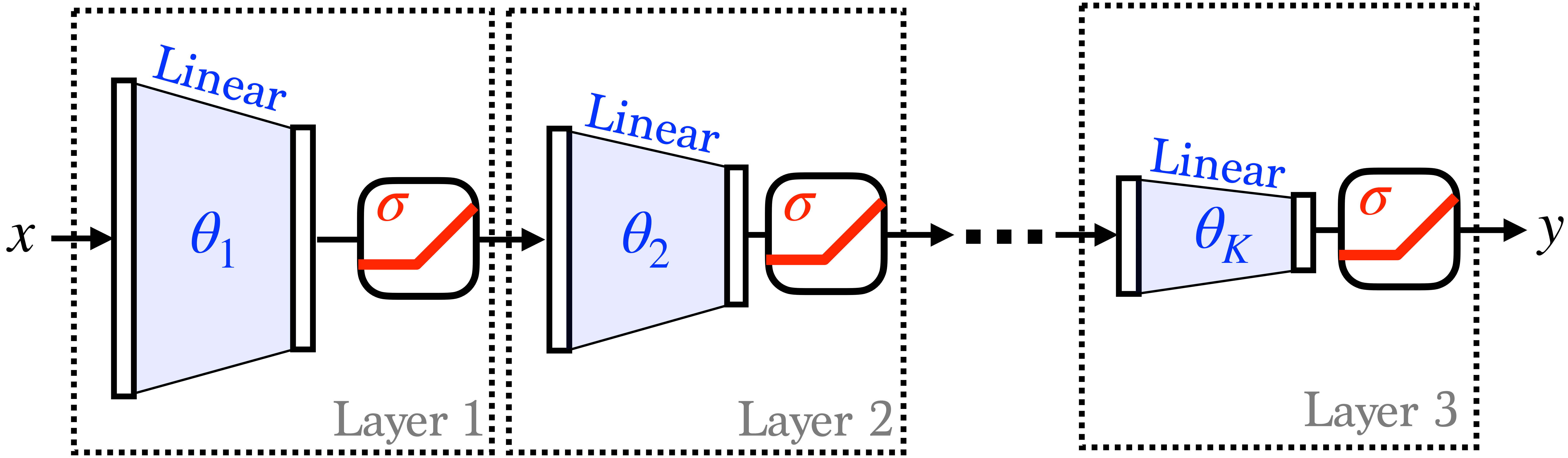


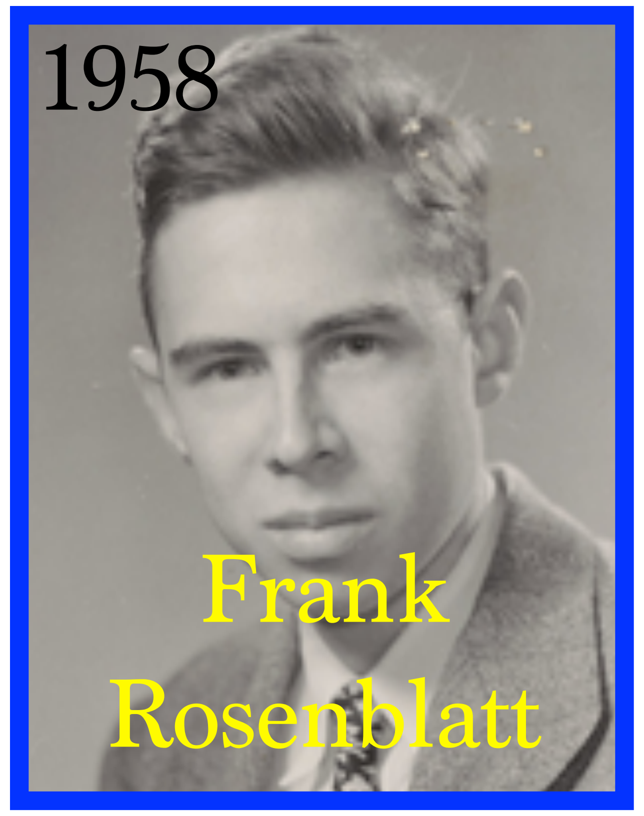




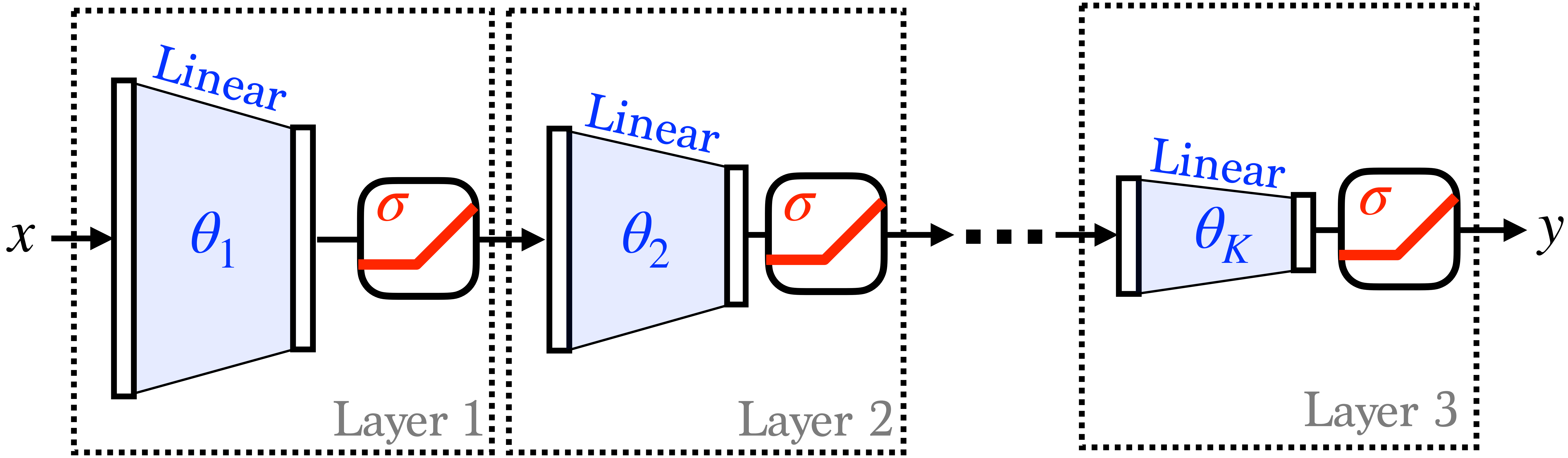


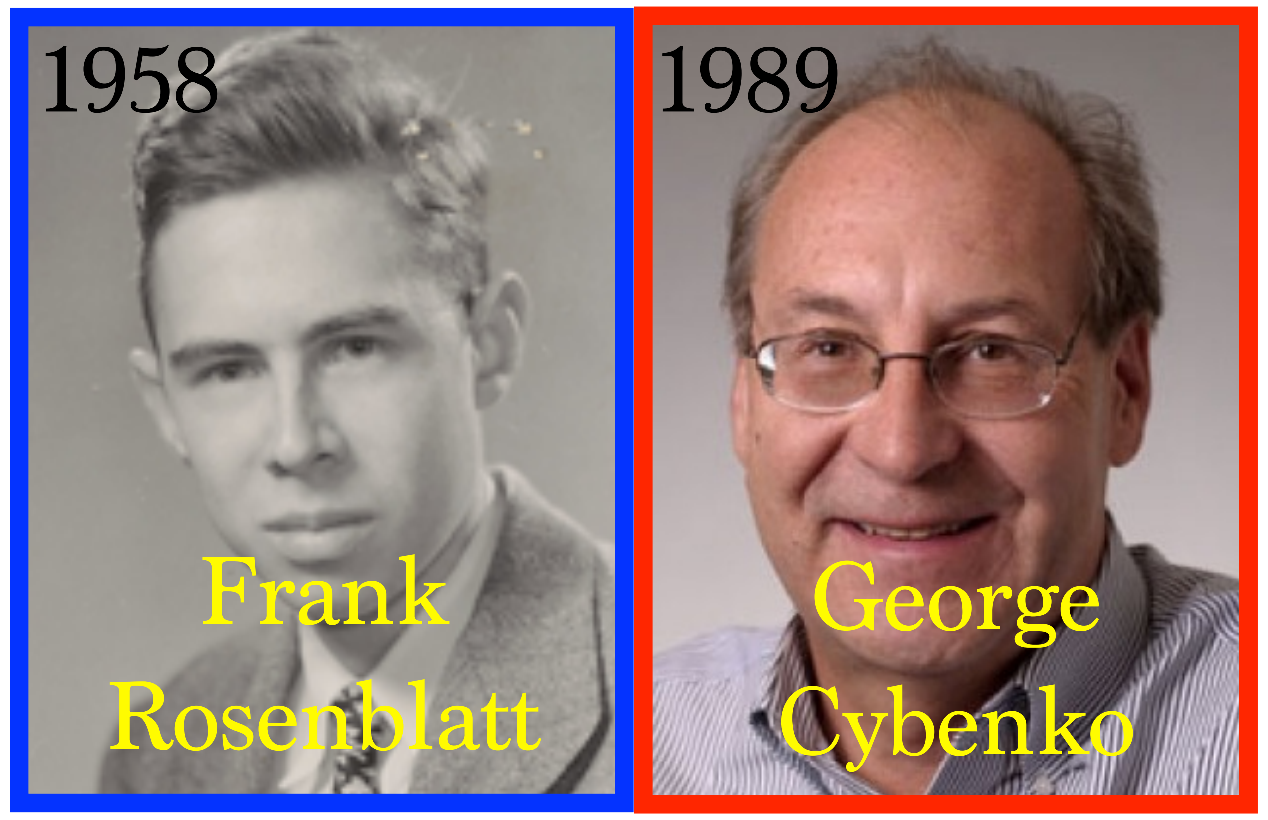
Perceptrons and Universality



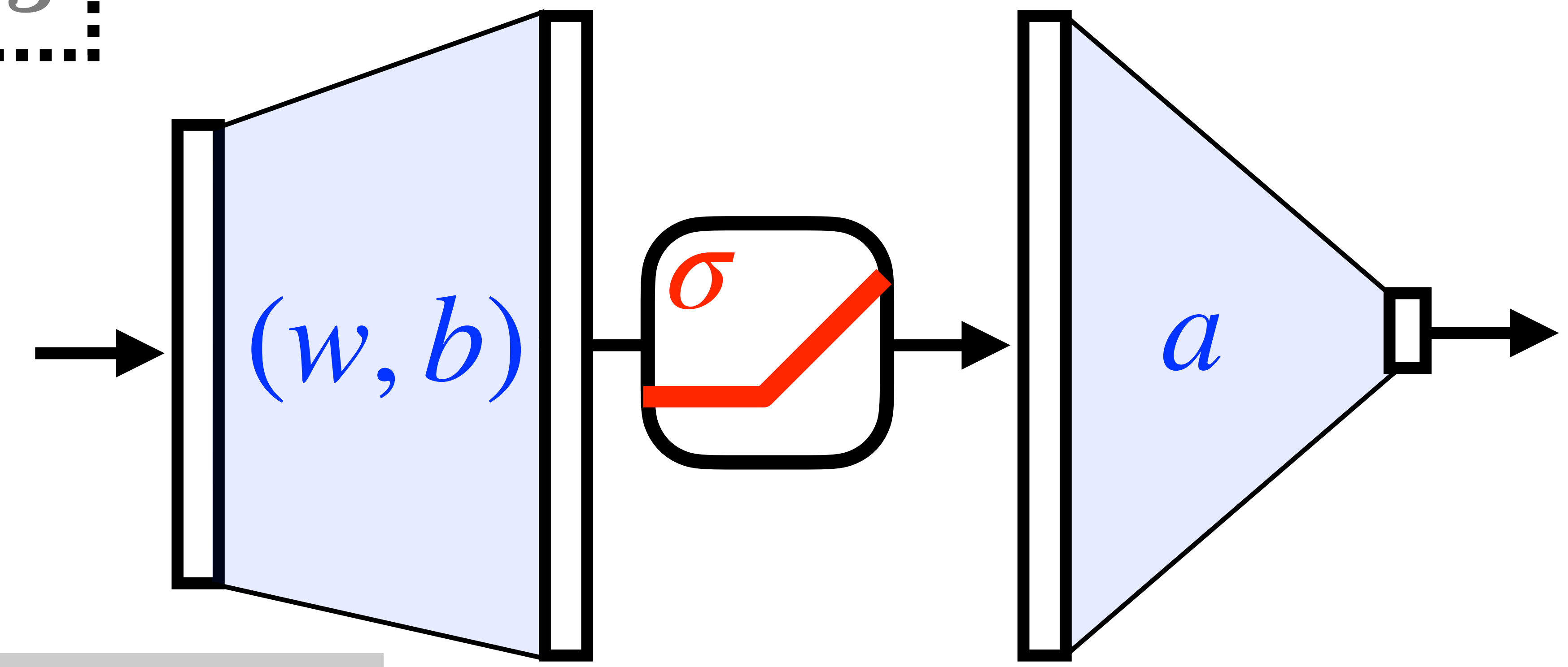


Perceptrons and Universality





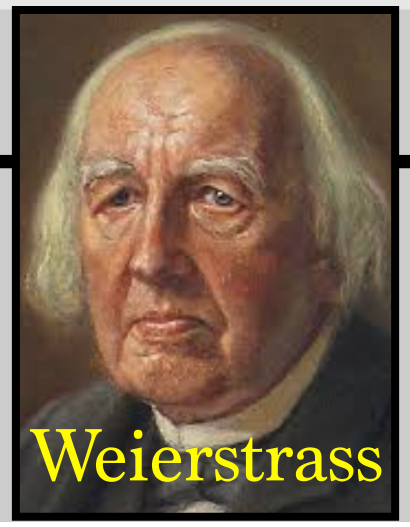
Two layers: $\Gamma_\theta(x) := \sum_{i=1}^n a_i \sigma(\langle x, w_i \rangle + b_i)$ $\qquad \theta = (a_i, w_i, b_i)_i$



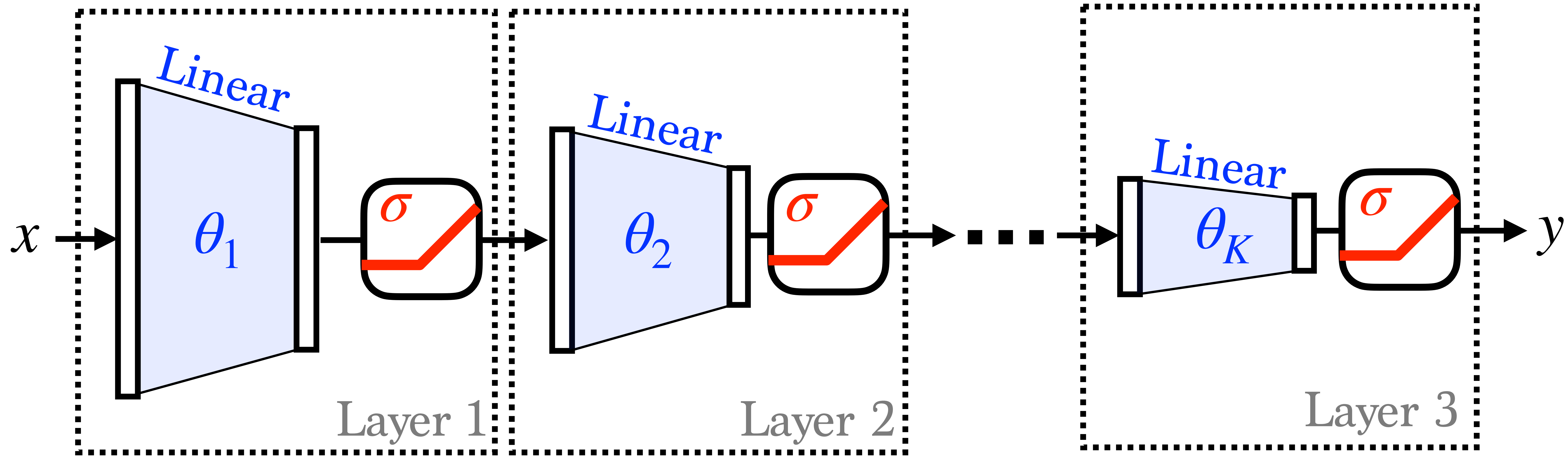
Theorem: for sigmoid and ReLu σ, on a compact Ω, $\{\Gamma_\theta\}_\theta$ is dense in $(\mathscr{C}(\Omega), L^\infty(\Omega))$.

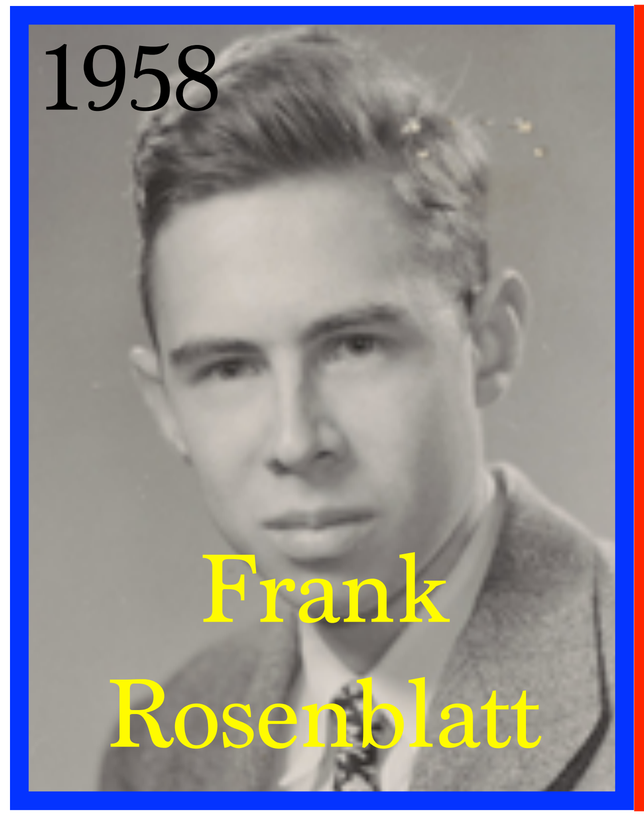
Proof: For $\sigma = \sin$, $\{\Gamma_\theta\}_\theta$ is an algebra $\longrightarrow$ $\{\Gamma_\theta\}_\theta$ is dense

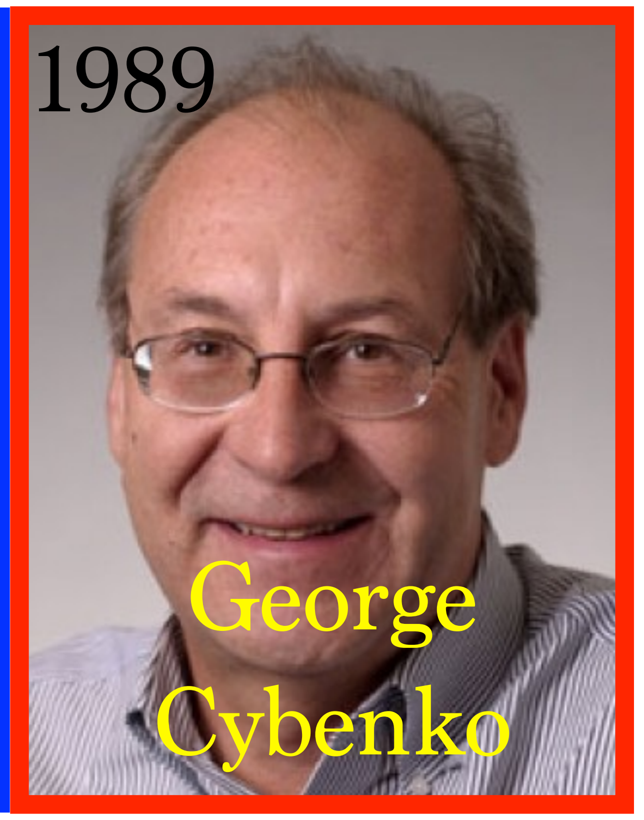
Approximate sin by generic σ



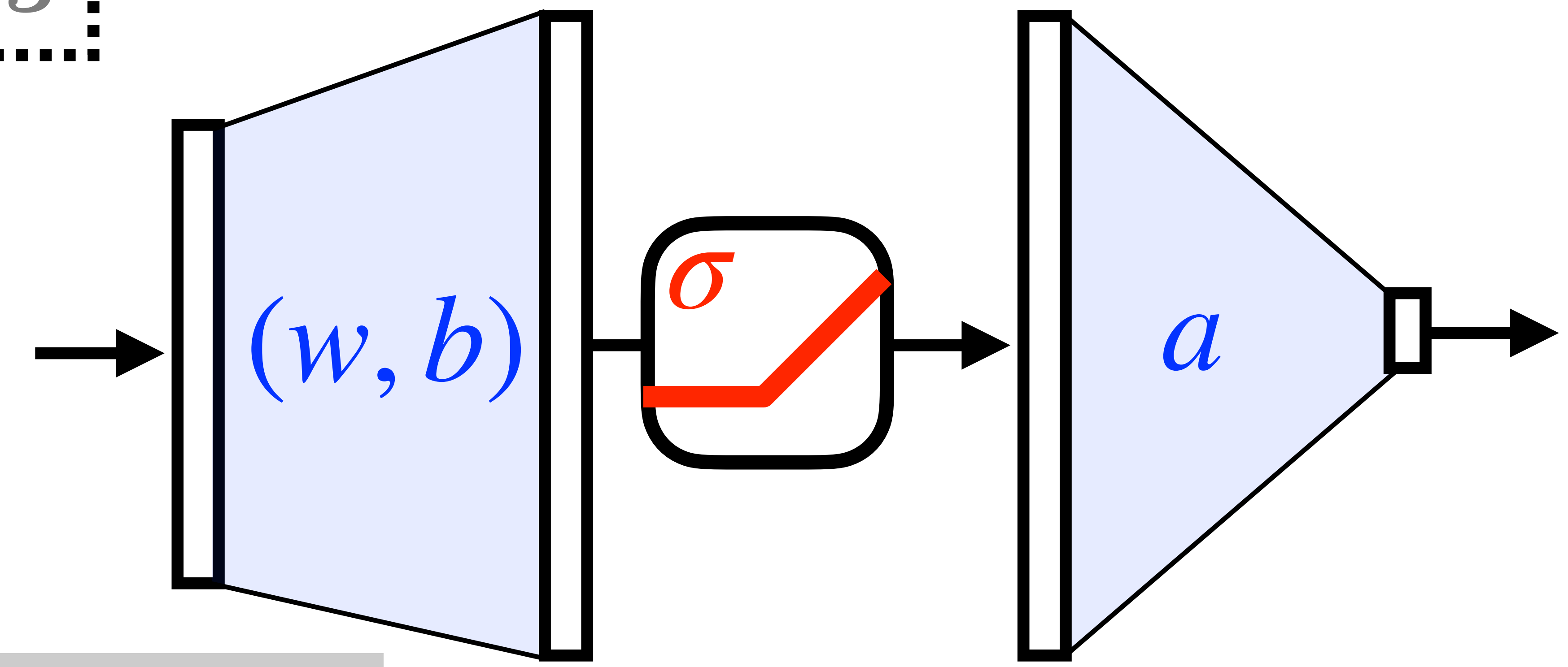
Perceptrons and Universality





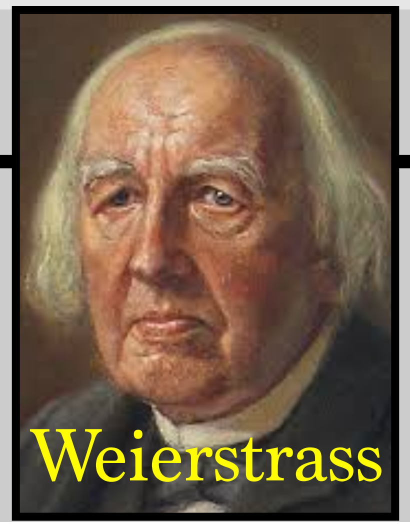


Two layers: $\Gamma_\theta(x) := \sum_{i=1}^{n} a_i \sigma(\langle x, w_i \rangle + b_i)$ $\qquad \theta = (a_i, w_i, b_i)_i$



Theorem: for sigmoid and ReLu σ, on a compact Ω, $\{\Gamma_\theta\}_\theta$ is dense in $(\mathscr{C}(\Omega), L^\infty(\Omega))$.

Proof: For $\sigma = \sin$, $\{\Gamma_\theta\}_\theta$ is an algebra $\longrightarrow$ $\{\Gamma_\theta\}_\theta$ is dense

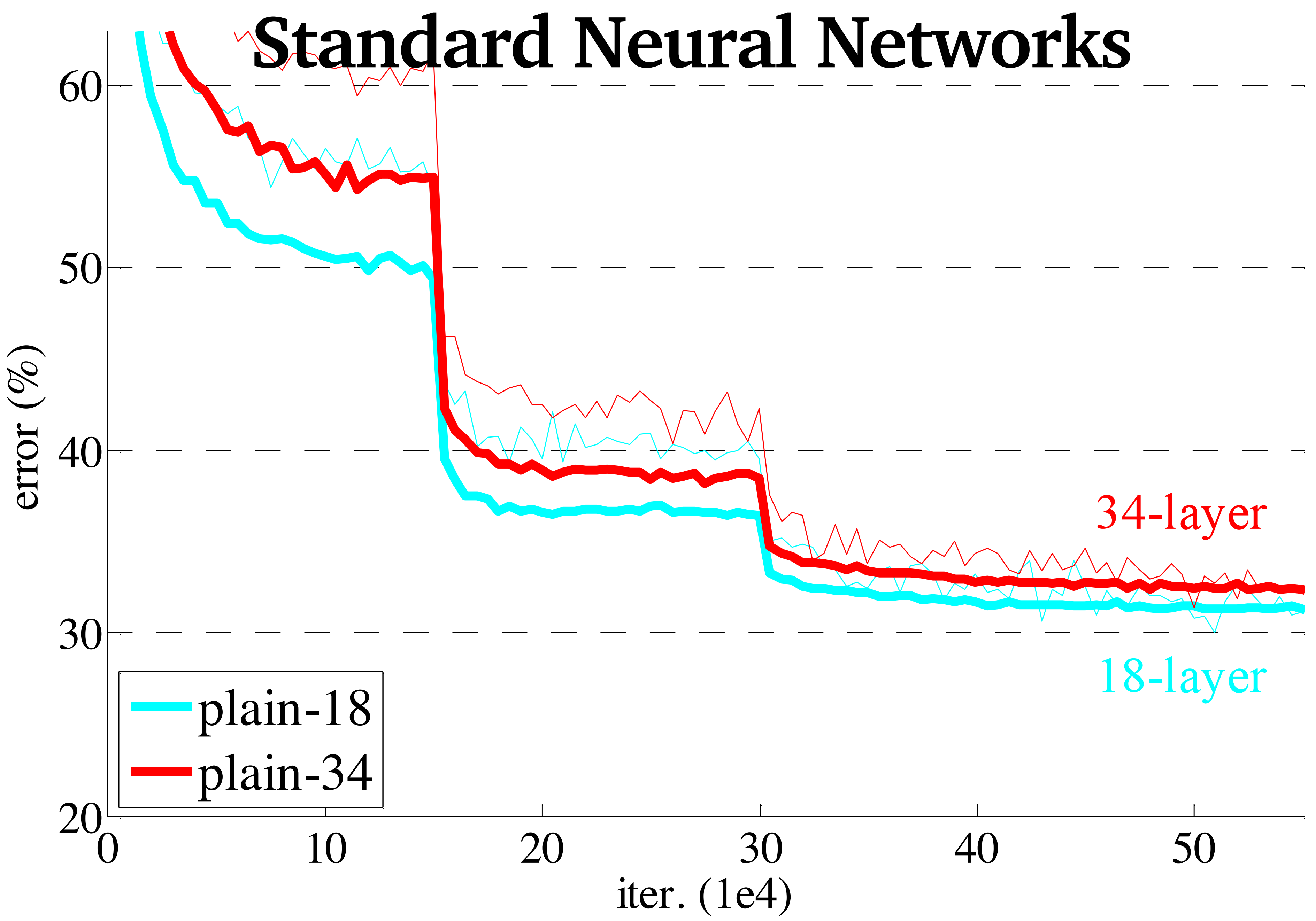


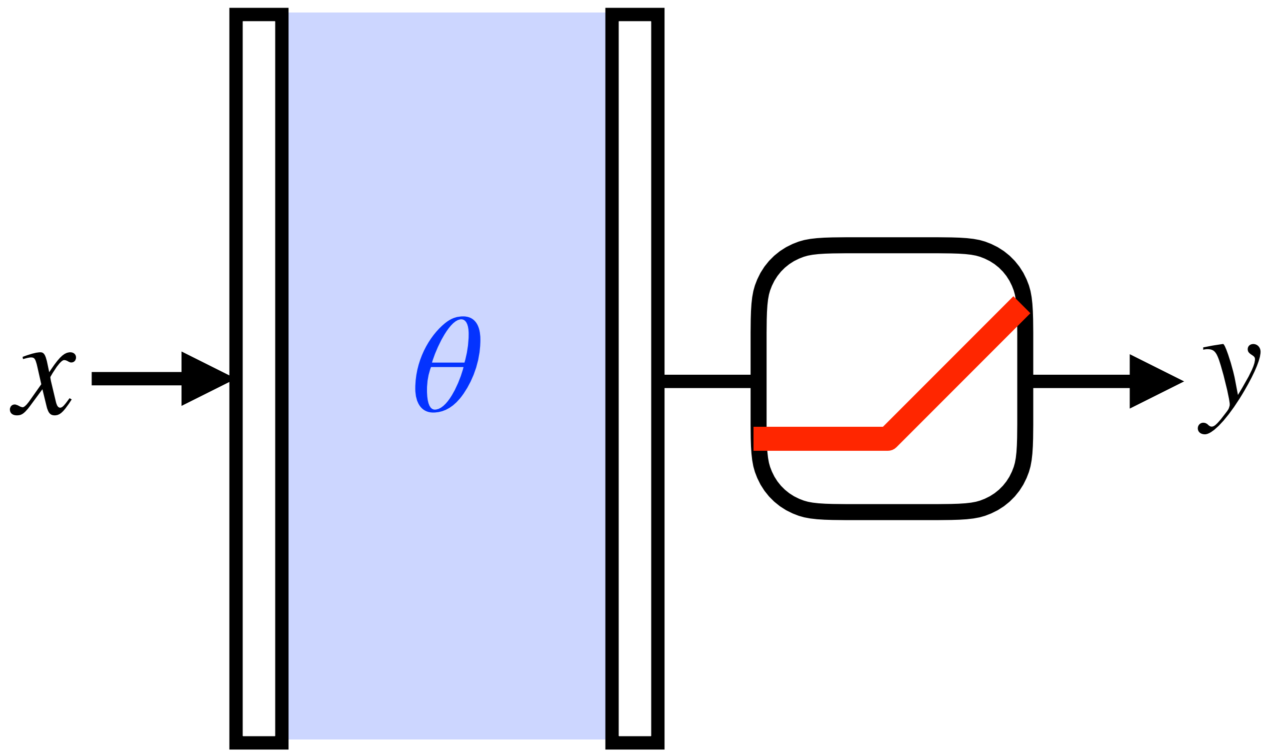
Approximate sin by generic σ

Open problems

- *Expressivity:* role of depth?
- *Optimization:* convergence of gradient descent.

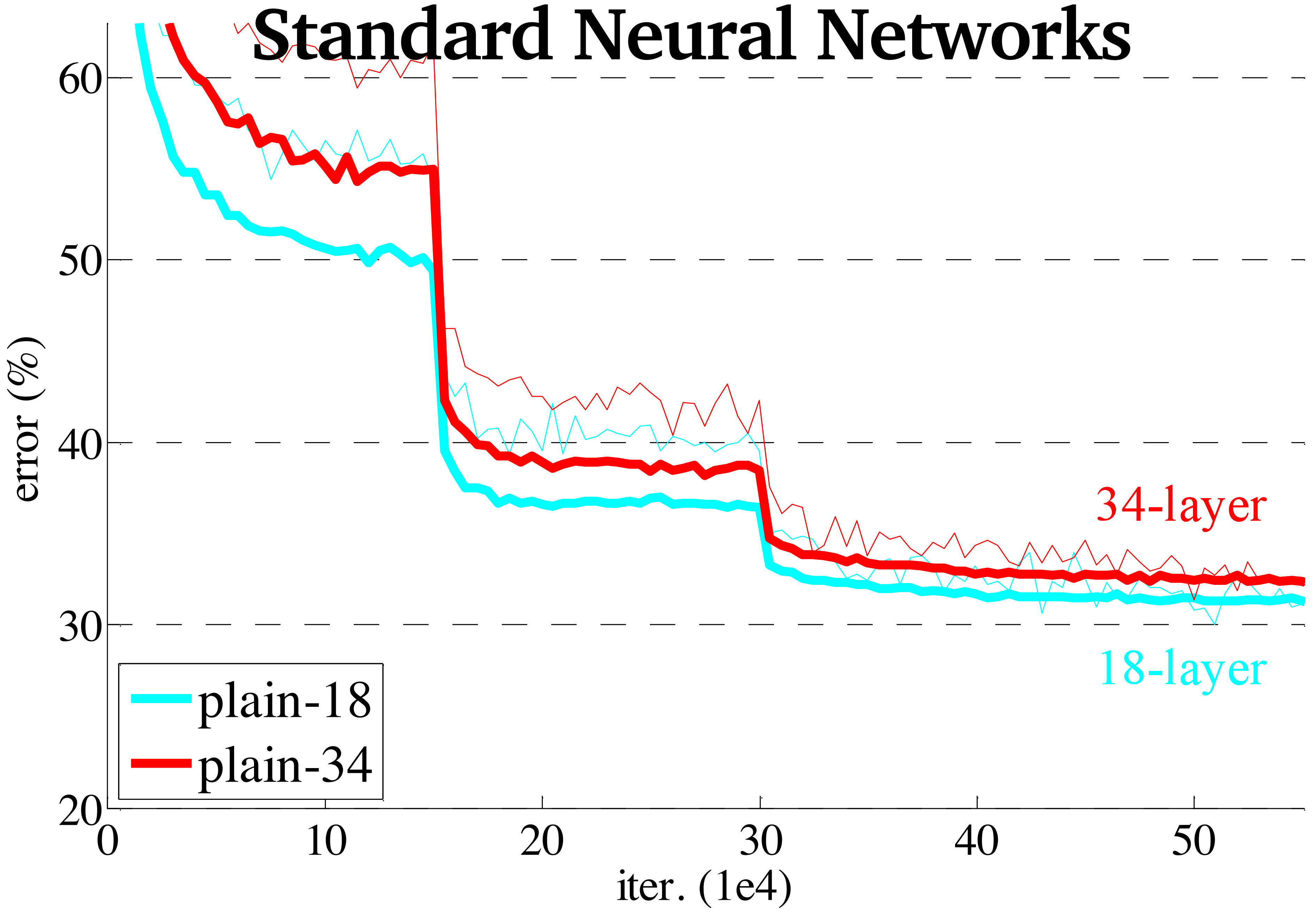
The Deeper, the Better

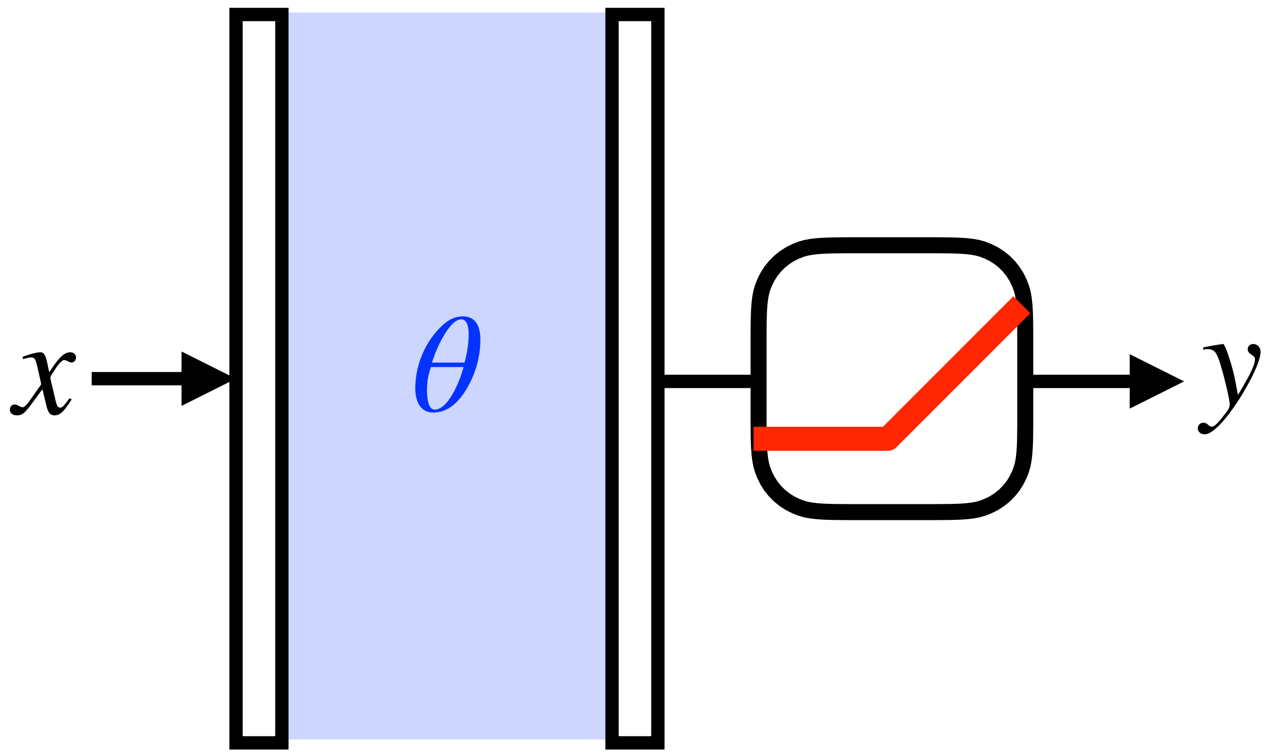




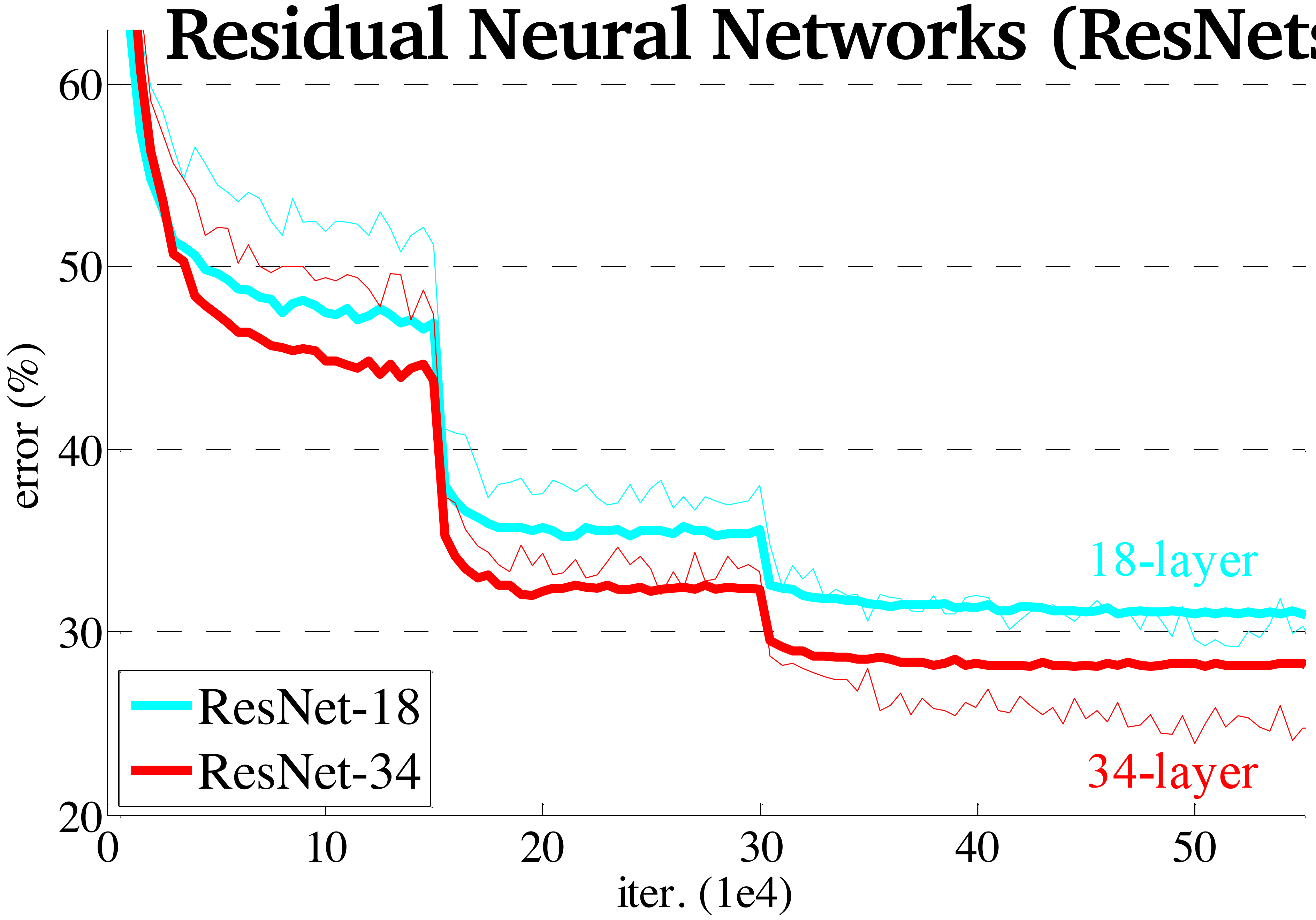
The Deeper, the Better

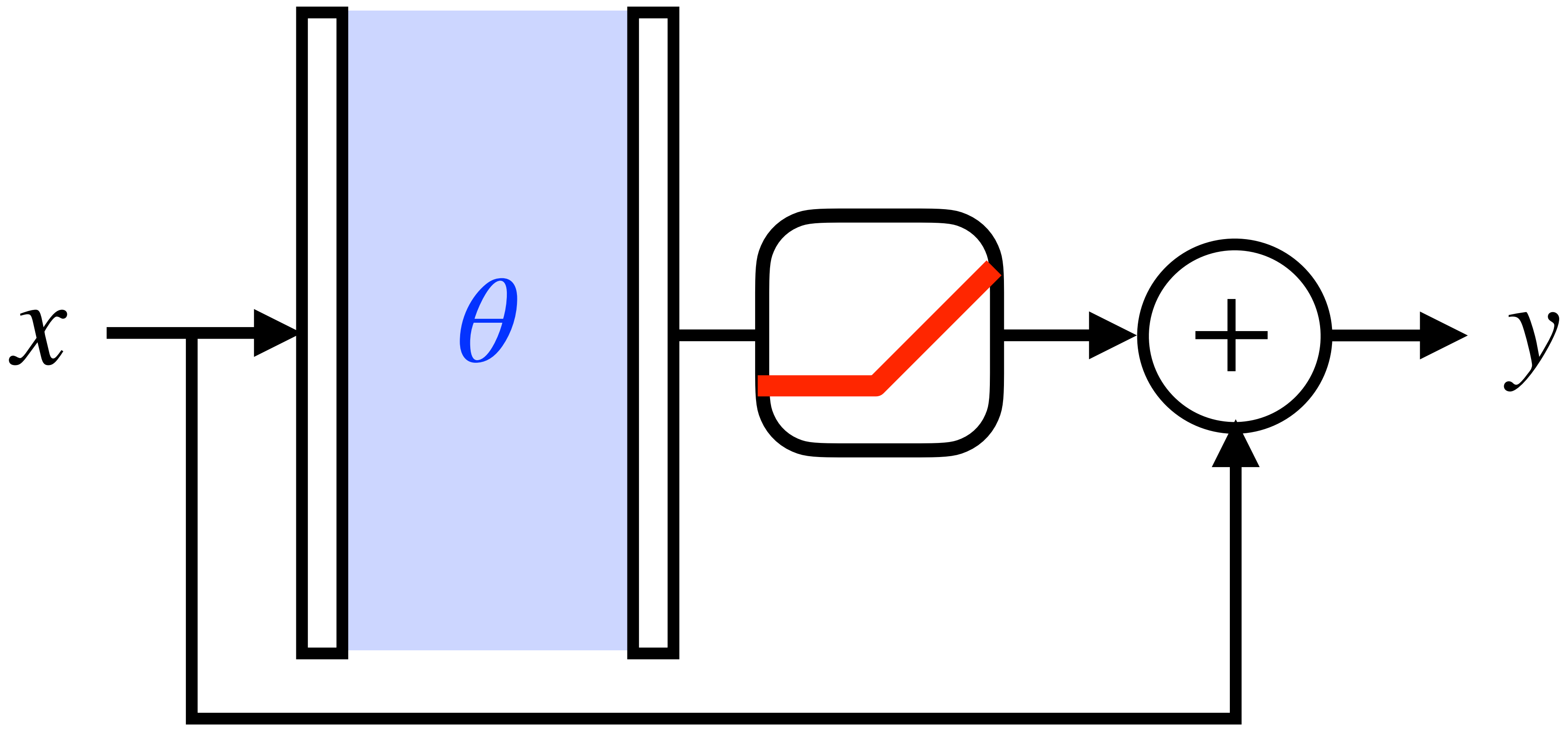
Standard Neural Networks





Residual Neural Networks (ResNets)





ResNets

Infinite depth

Differential equations

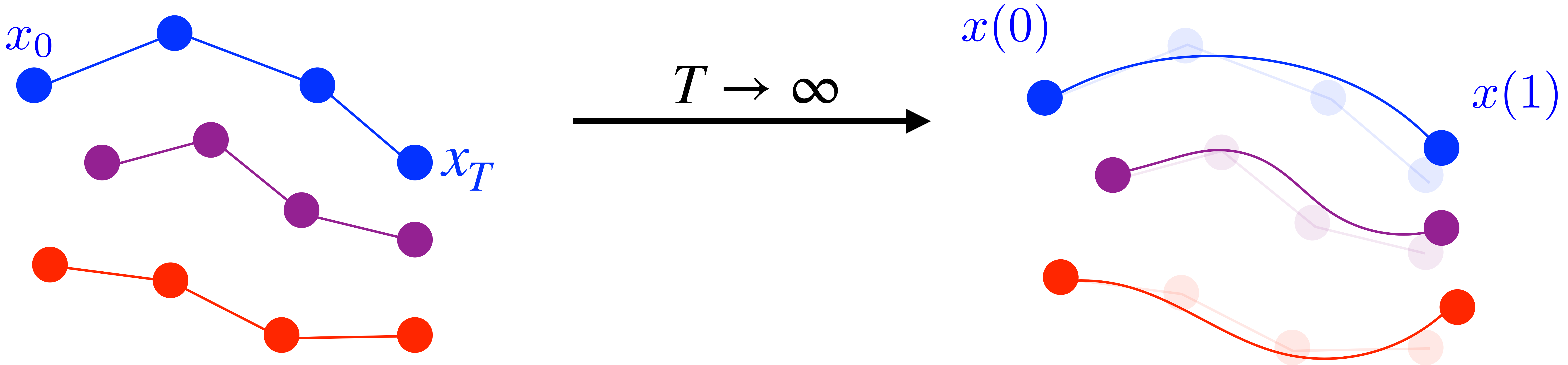
Infinite Depth and Neural-ODEs

T layers

$$x_{t+1} = x_t + \frac{1}{T}\Gamma_{\theta_t}(x_t)$$

$T = \infty$ layers

$$\dot{x}_t = \Gamma_{\theta(t)}(x(t))$$



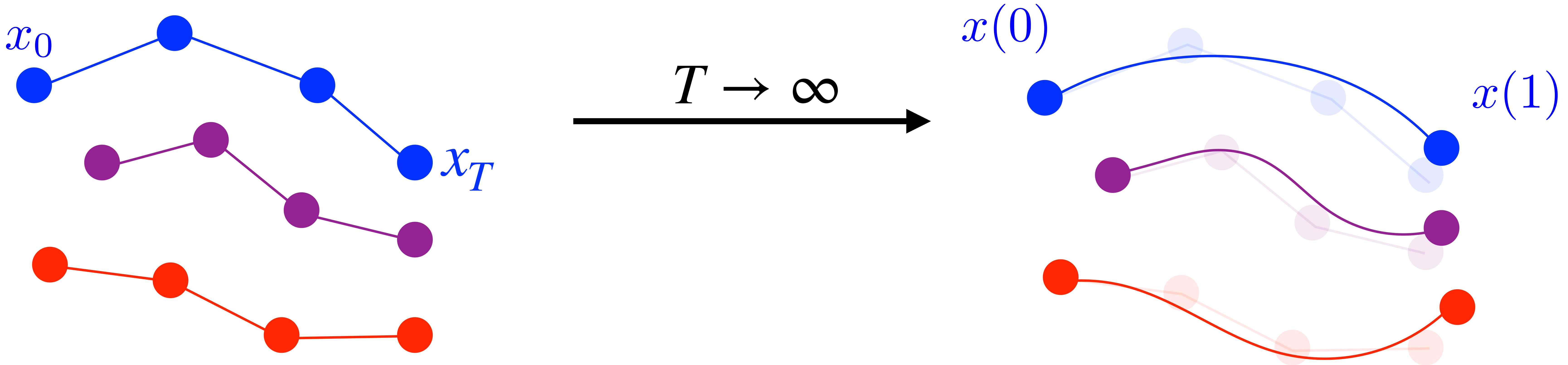
Infinite Depth and Neural-ODEs

T layers

$$x_{t+1} = x_t + \frac{1}{T}\Gamma_{\theta_t}(x_t)$$

$T = \infty$ layers

$$\dot{x}_t = \Gamma_{\theta(t)}(x(t))$$



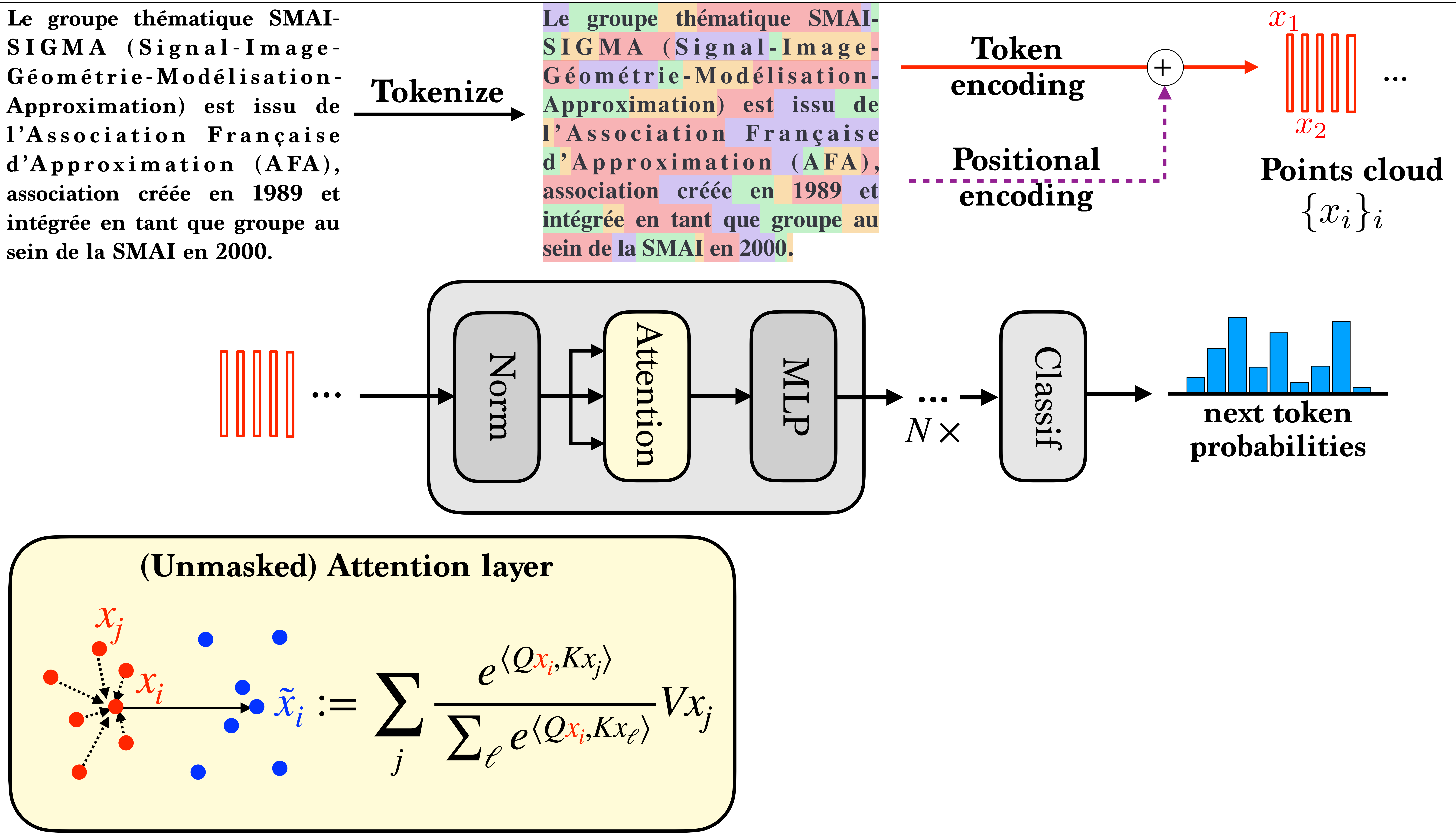
Trajectories cannot cross: $x \to x(1)$ defines a diffeomorphism.

Open problems

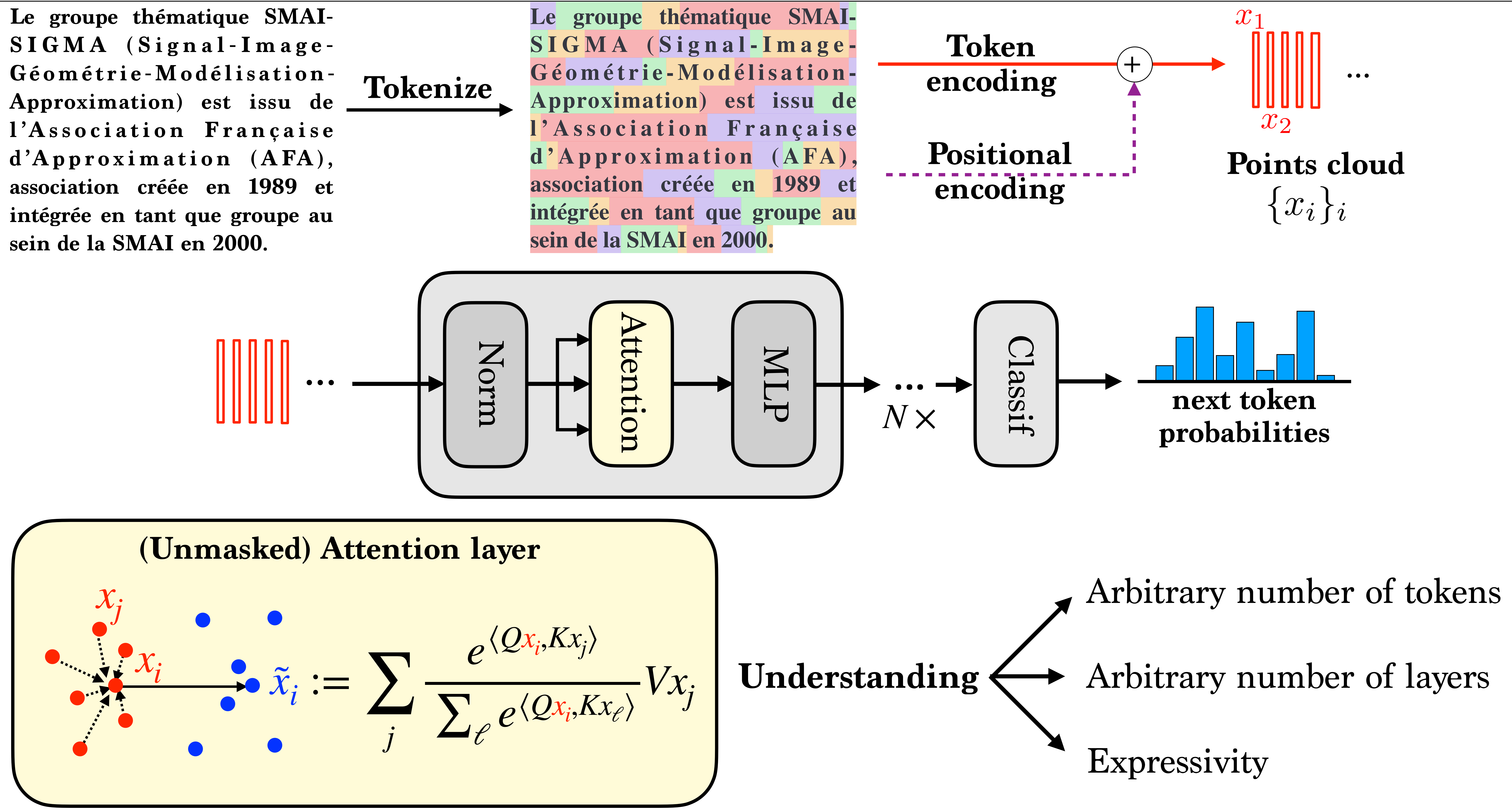
- *Expressivity:* universality requires embedding in $\mathbb{R}^{d+1}$ space.
- *Optimization:* weights could « blows » during training.


In Context Mappings

Transformers and attention mechanism



Transformers and attention mechanism

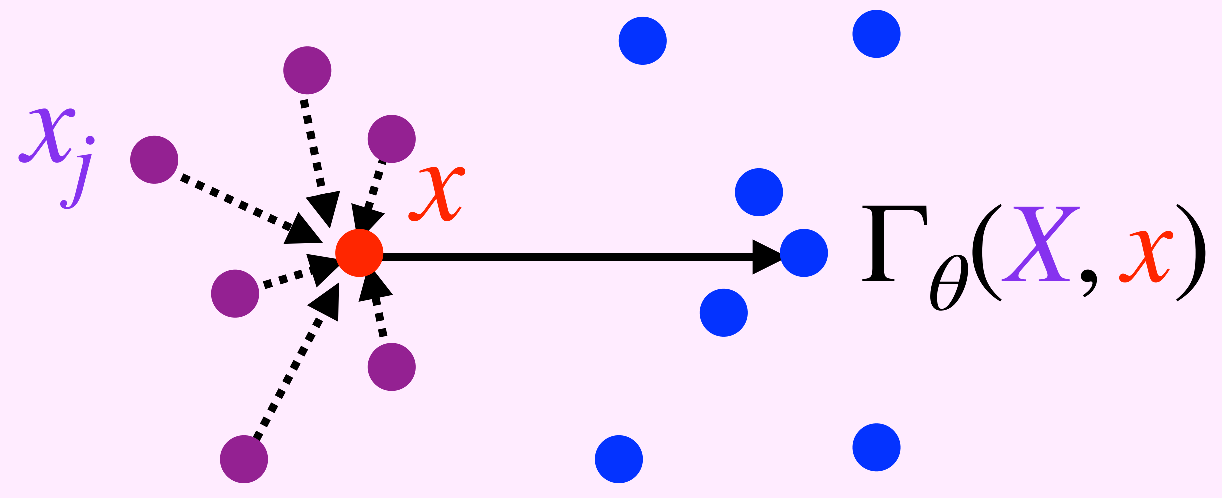


Attention as In-context Mapping

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

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

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

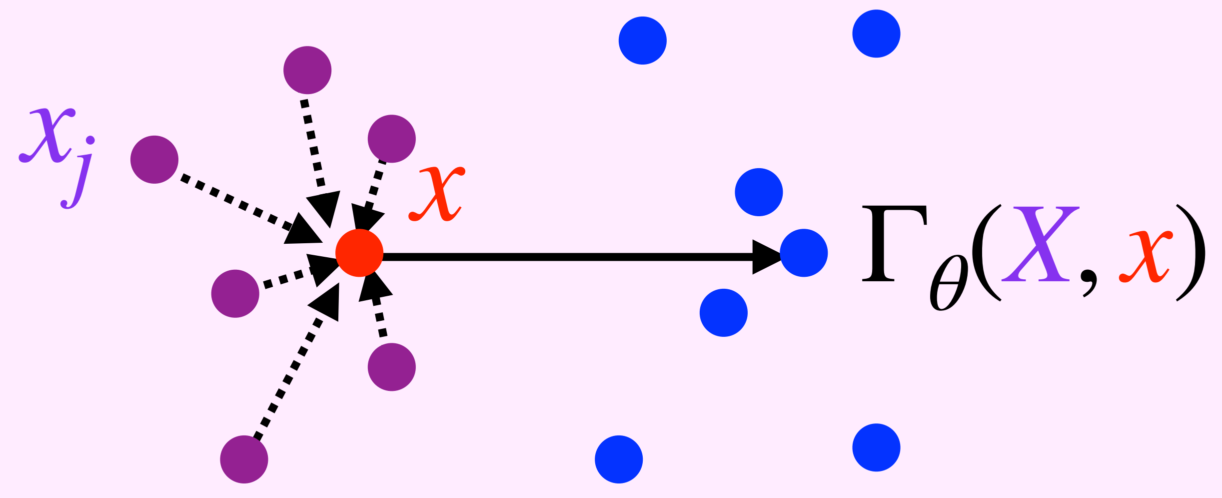


Attention as In-context Mapping

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: $X \mapsto \left\{\sum_{h=1}^H \Gamma_{\theta_h}[X](x_i)\right\}_{i=1}^n$

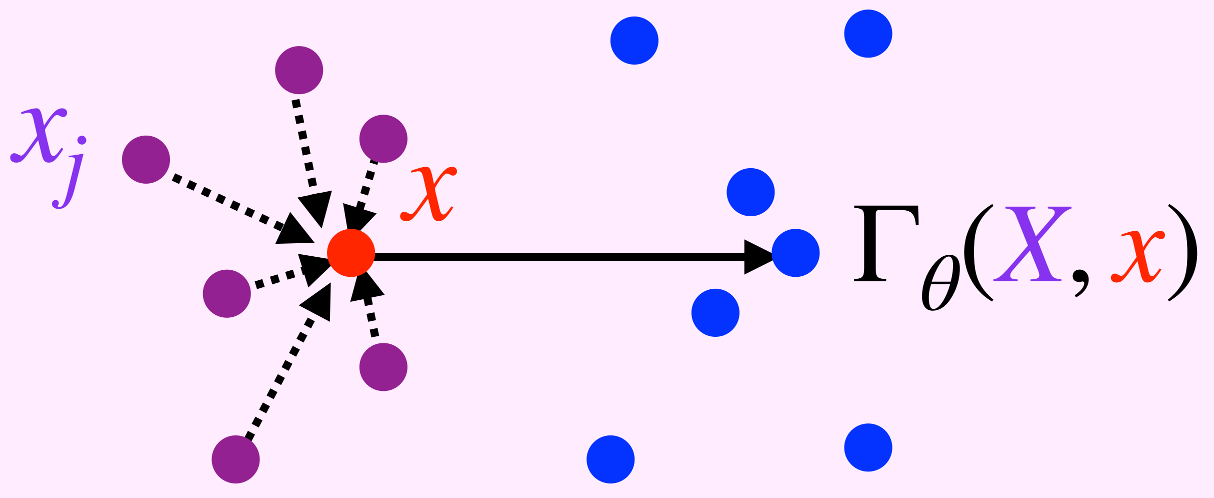
K_1, Q_1
K_2, Q_2
...
K_H, Q_H

Attention as In-context Mapping

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

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

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



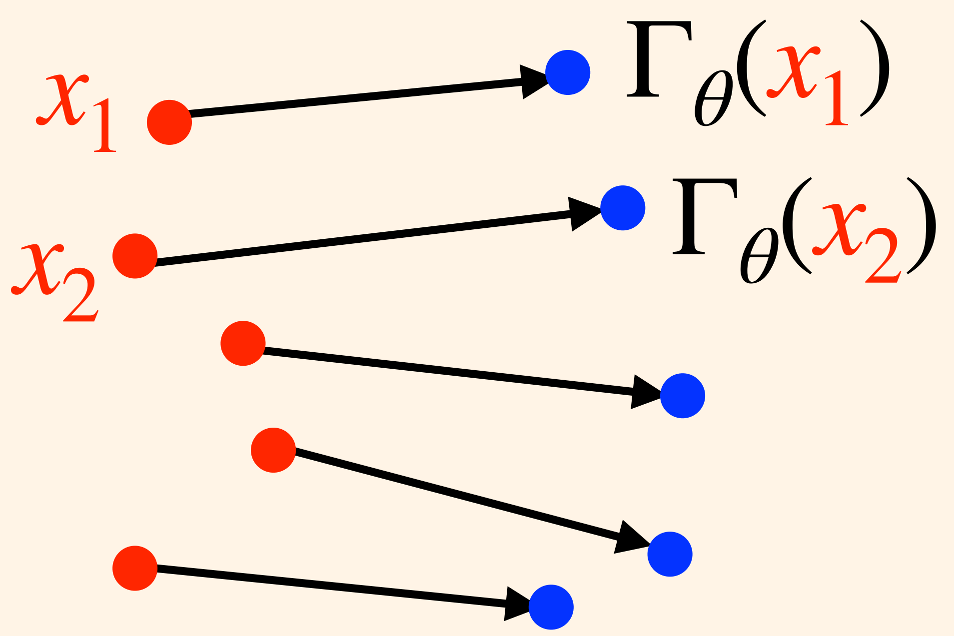
Single-head attention layer: $X \mapsto \left\{\Gamma_\theta[X](x_i)\right\}_{i=1}^n$

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

K_1, Q_1
K_2, Q_2
...
K_H, Q_H

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

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

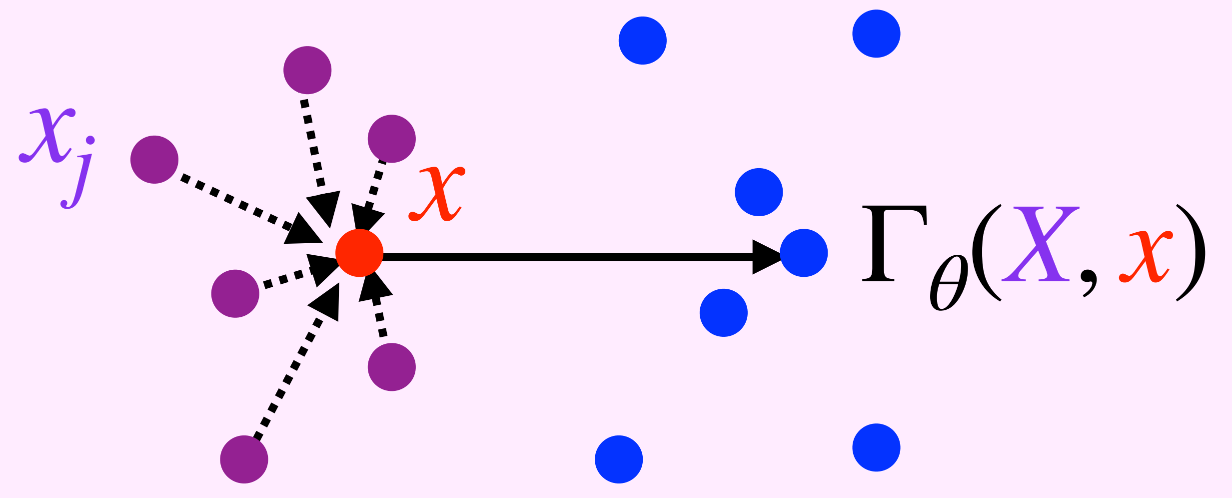


Attention as In-context Mapping

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

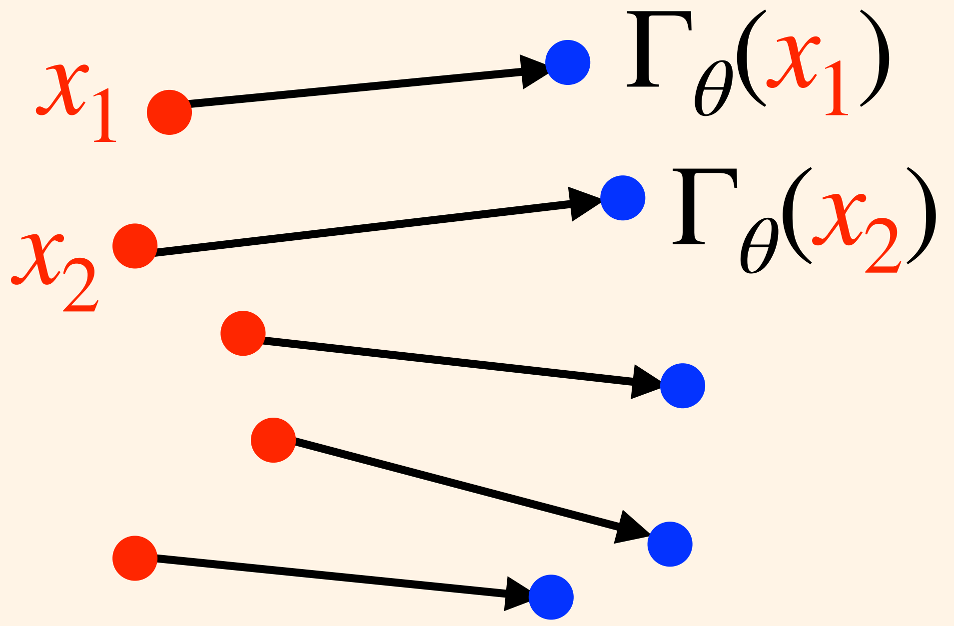
Multi-head attention layer: $X \mapsto \{\sum_{h=1}^H \Gamma_{\theta_h}[X](x_i)\}_{i=1}^n$

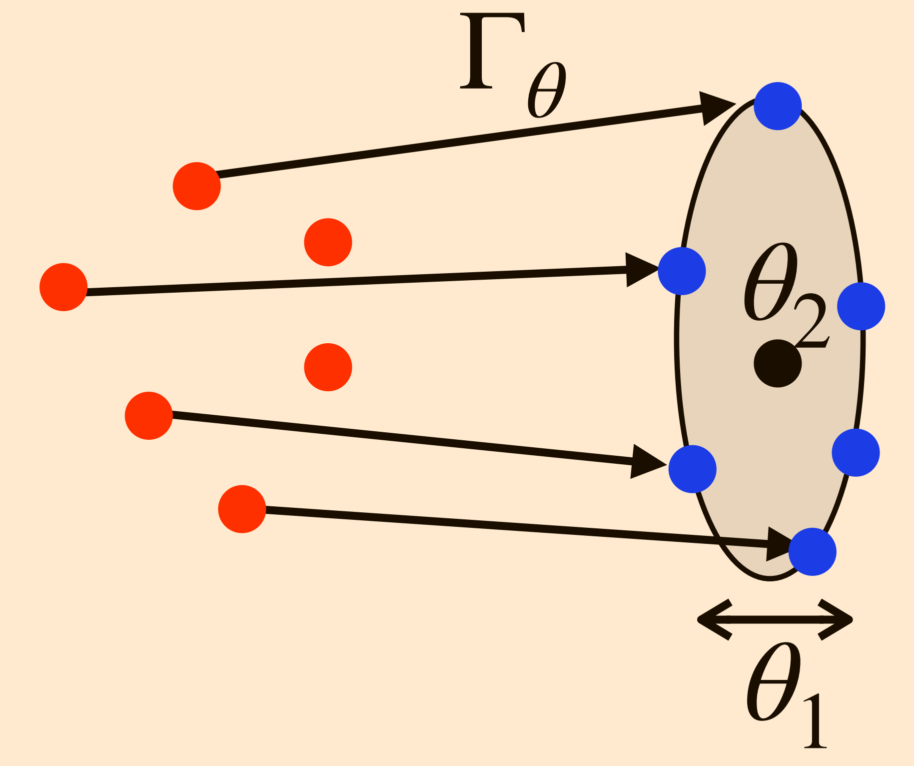
K_1, Q_1
K_2, Q_2
...
K_H, Q_H

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

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

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



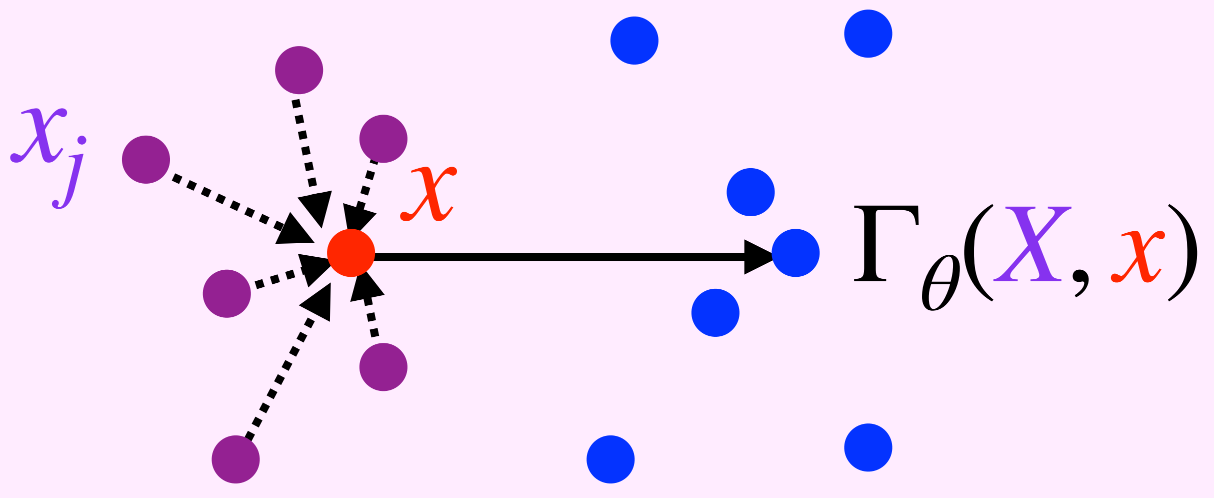


Attention as In-context Mapping

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

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

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

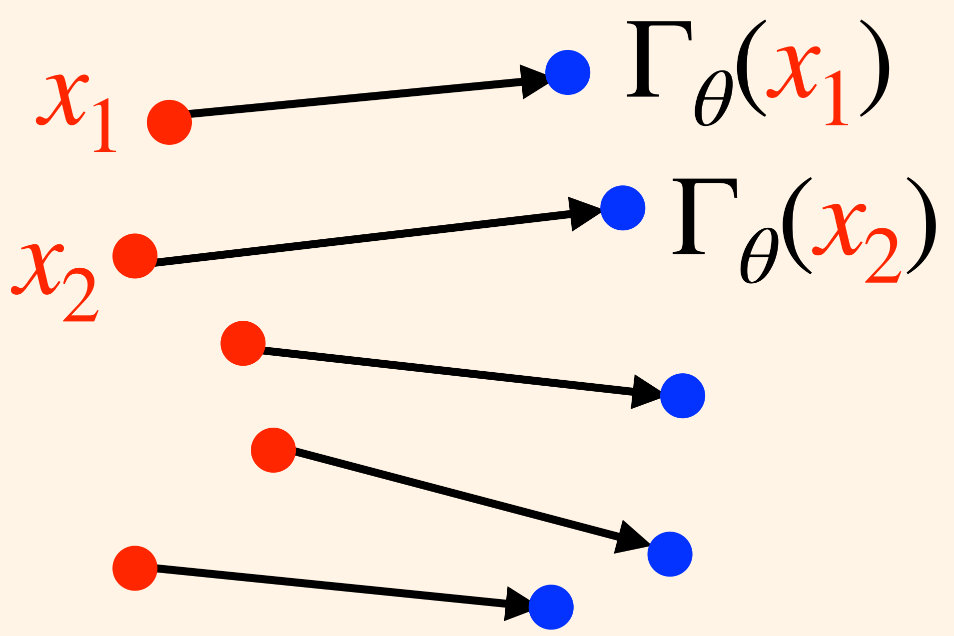


Single-head attention layer: $X \mapsto \left\{\Gamma_\theta[X](x_i)\right\}_{i=1}^n$

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

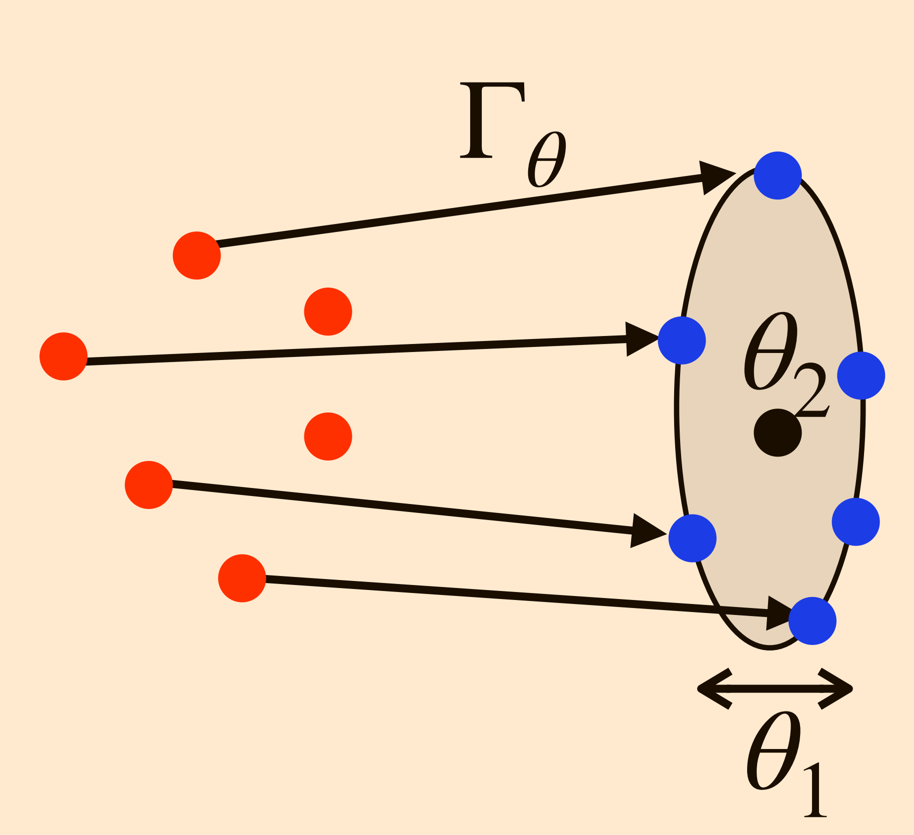
K_1, Q_1
K_2, Q_2
...
K_H, Q_H

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

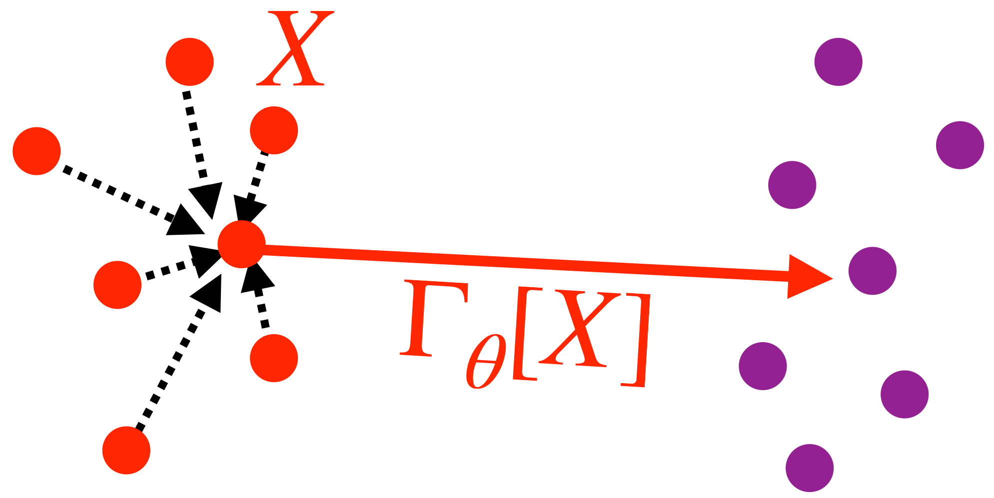


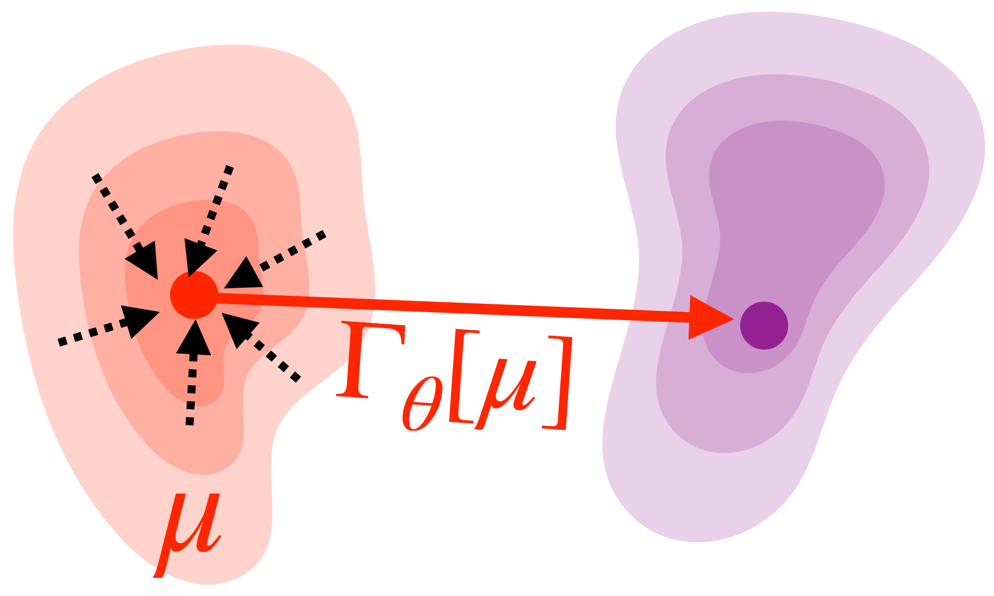
Transformer ≡ composition of in-context and context-free layers.

Attentions Operating over Measures

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

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



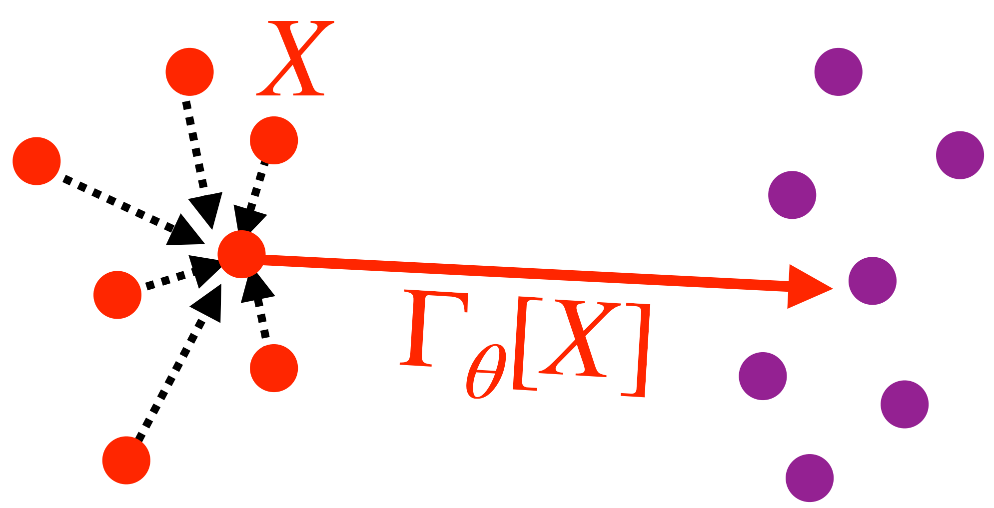


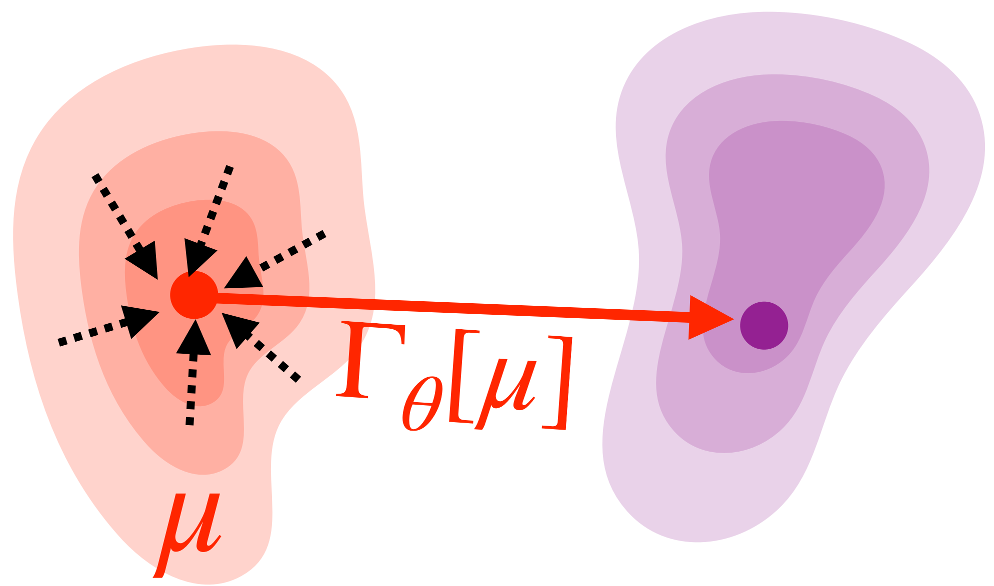
Attentions Operating over Measures

Number n of token is arbitrary.

(Unmasked) attention is permutation invariant.

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



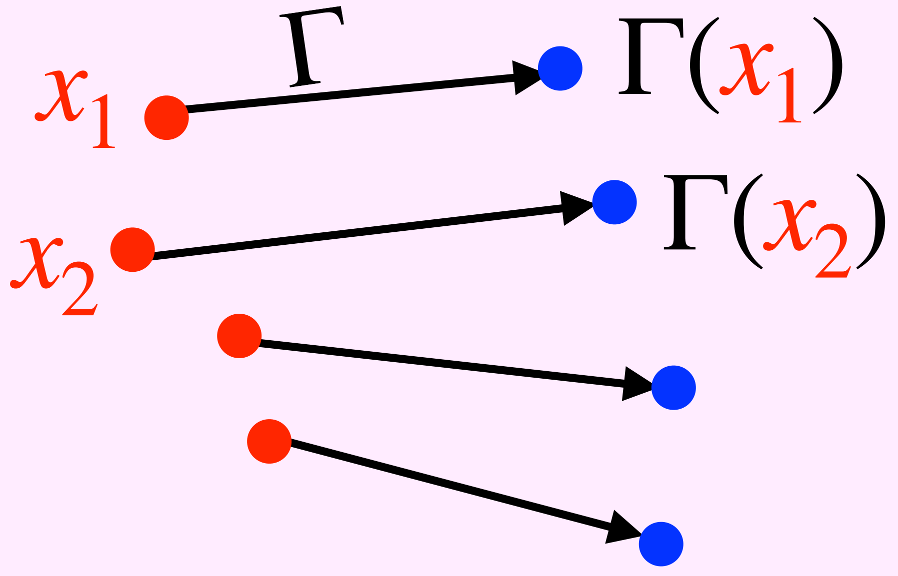


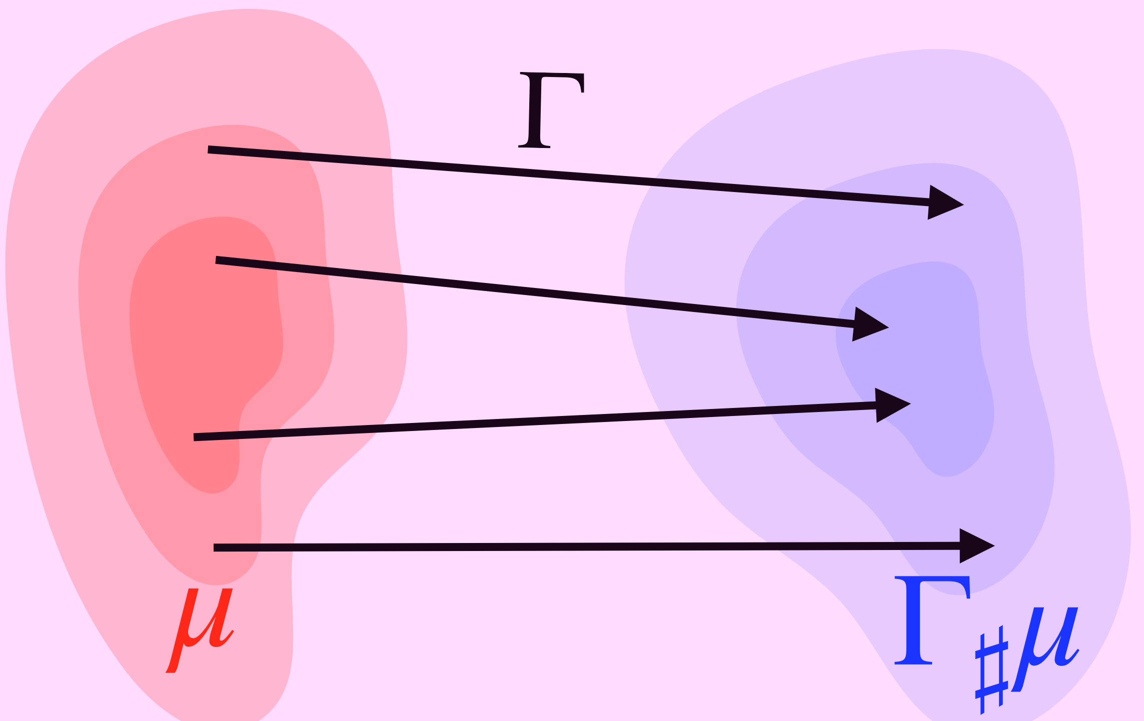
Push-forward

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

$$\downarrow$$

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





Layers

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

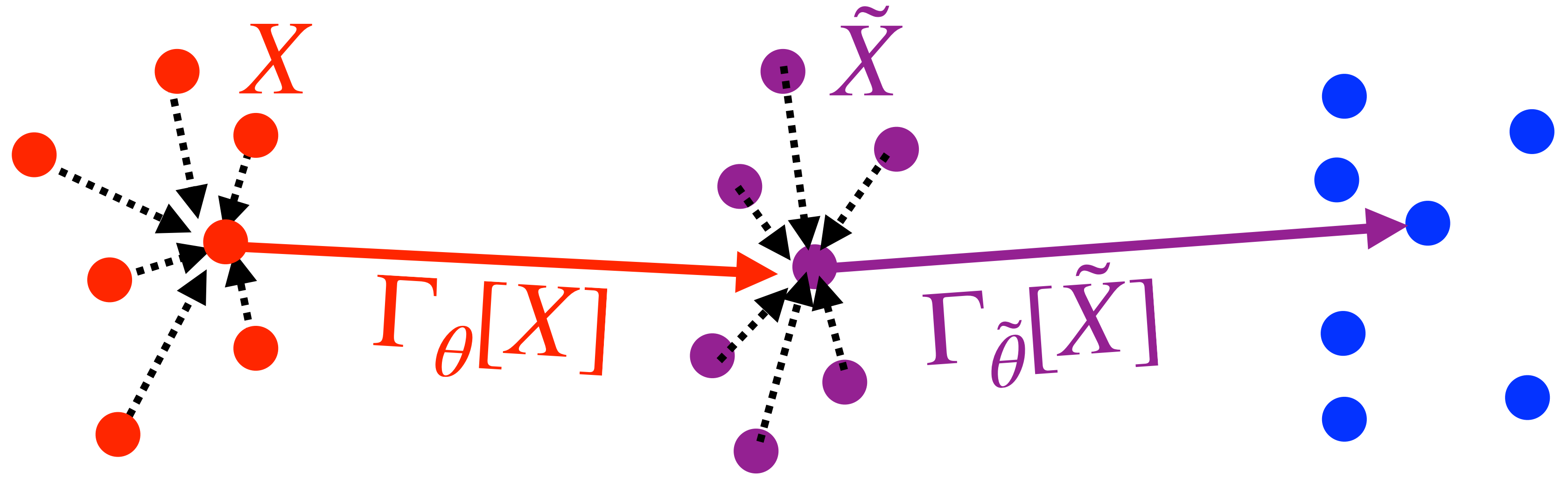
$$\downarrow$$

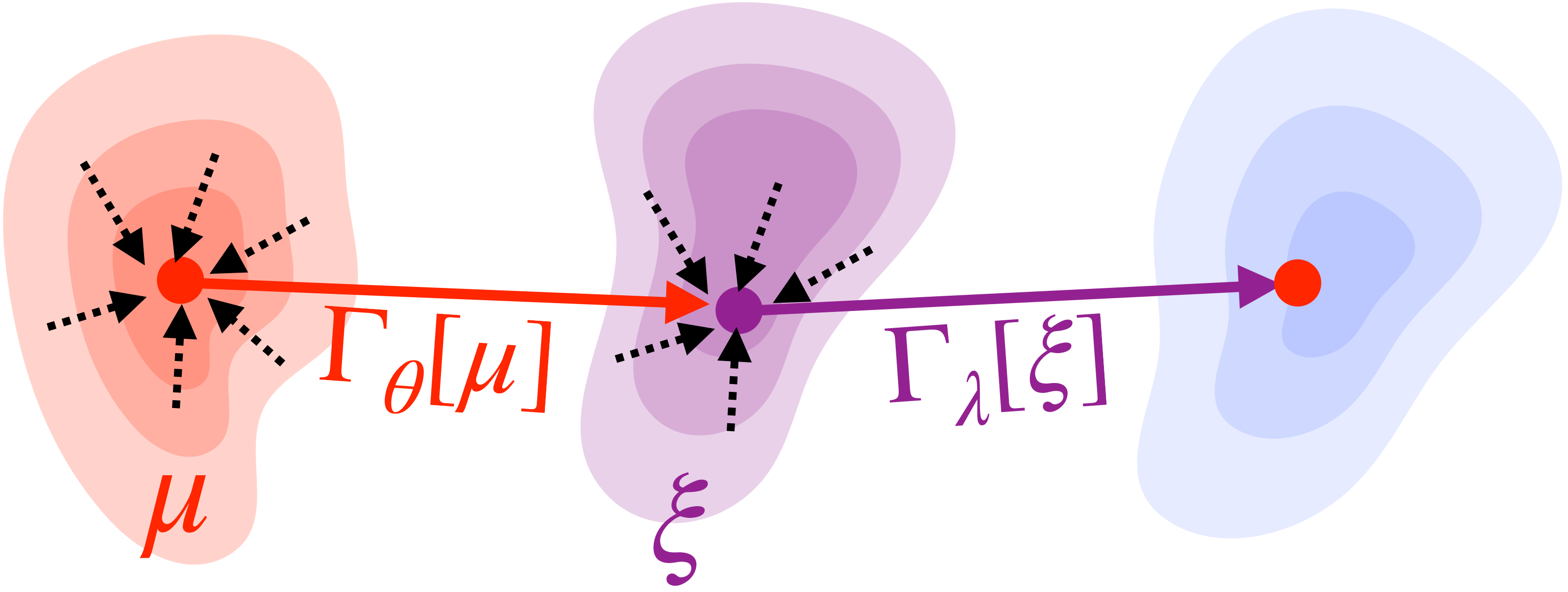
$$\mu \mapsto \Gamma[\mu]_\sharp \mu$$

Attentions Operating over Measures

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

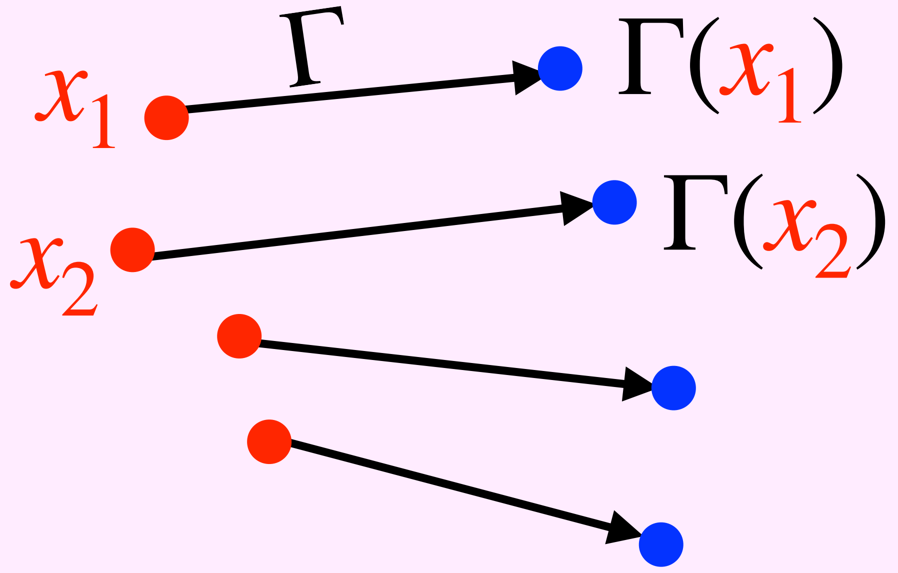
$$\Gamma_\theta[X](x) := \sum_j \frac{e^{\langle Qx, Kx_j\rangle}}{\sum_\ell e^{\langle Qx, Kx_\ell\rangle}} Vx_j \longrightarrow \boxed{\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}} \longrightarrow \Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle}\mathrm{d}\mu(y)}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



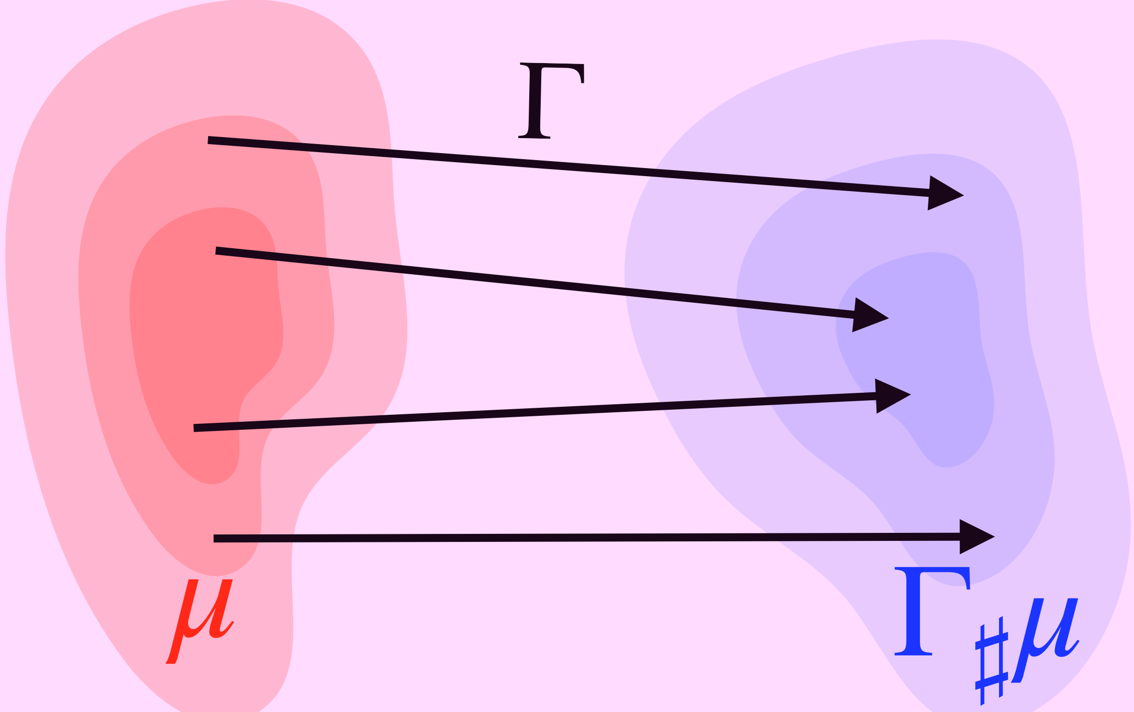


Push-forward

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



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



Layers

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

$$\mu \mapsto \Gamma[\mu]_\sharp \mu$$

Composing layers

$$(\Gamma' \diamond \Gamma)[X] := \Gamma'[Y] \circ \Gamma[X]$$

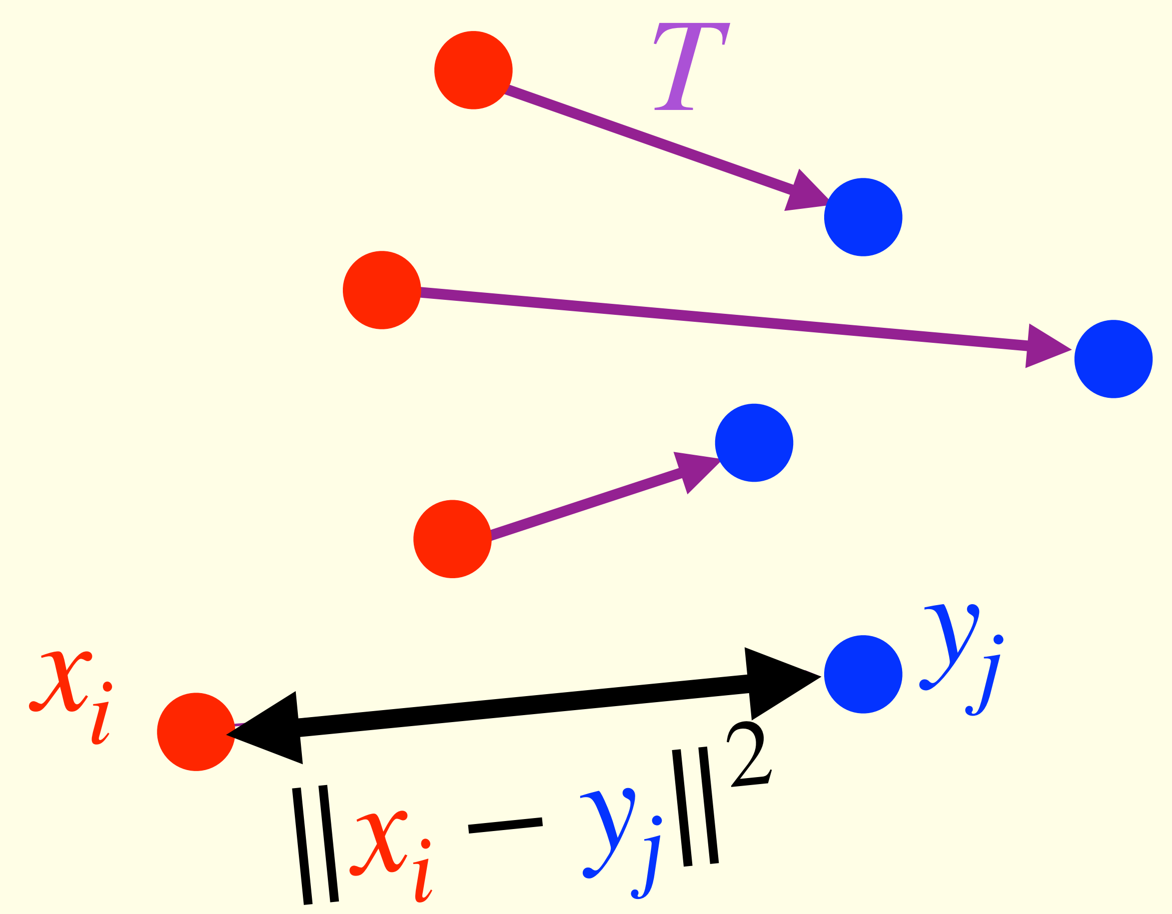
where $Y := (\Gamma[X](x_i))_i$

$$(\Gamma' \diamond \Gamma)[\mu] := \Gamma'[\xi] \circ \Gamma[\mu]$$

where $\xi := \Gamma[\mu]_\sharp \mu$


Transformers:
in-context universality

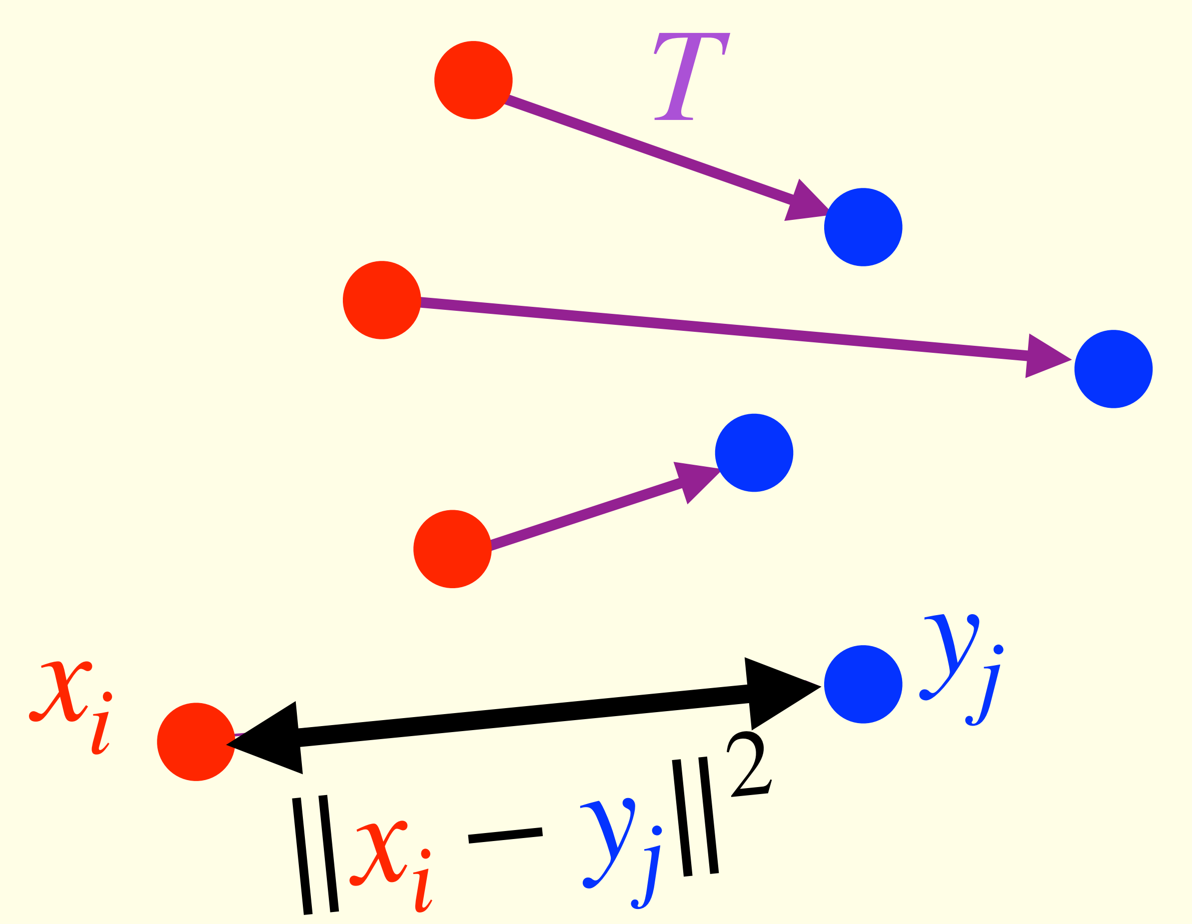
Optimal Transport (Wasserstein) Distance

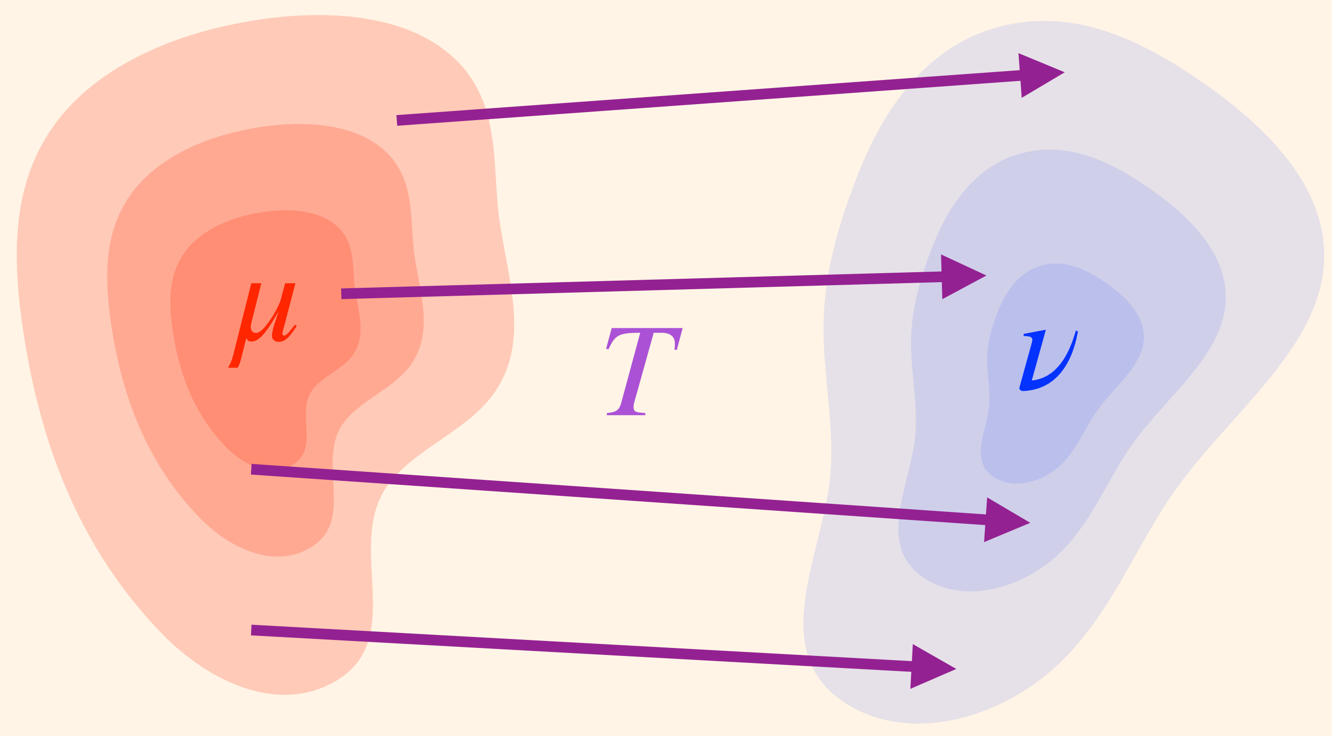


$$W_2(\mu, \nu)^2 := \min_{T} \sum_{i=1}^{n} \|x_i - y_{T(i)}\|^2$$

Monge 1784

Optimal Transport (Wasserstein) Distance

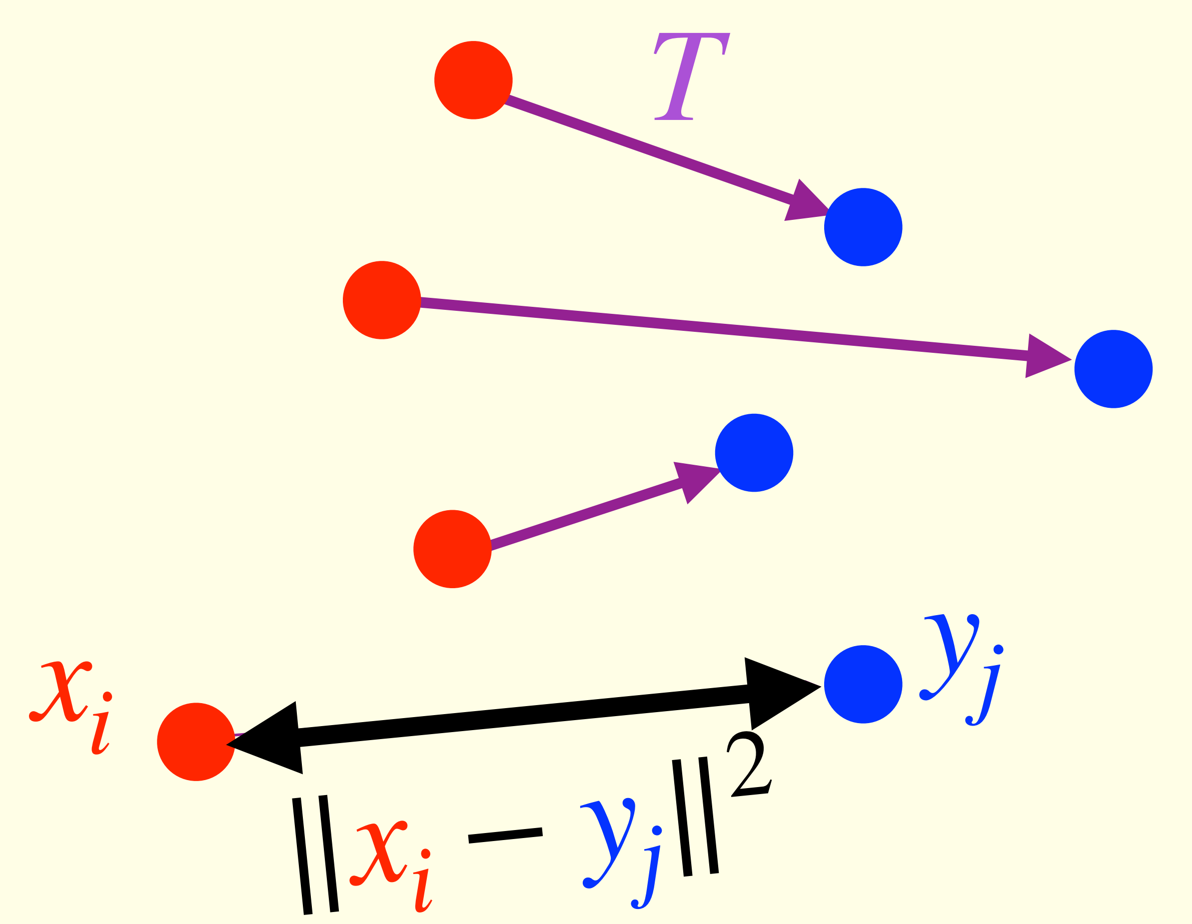


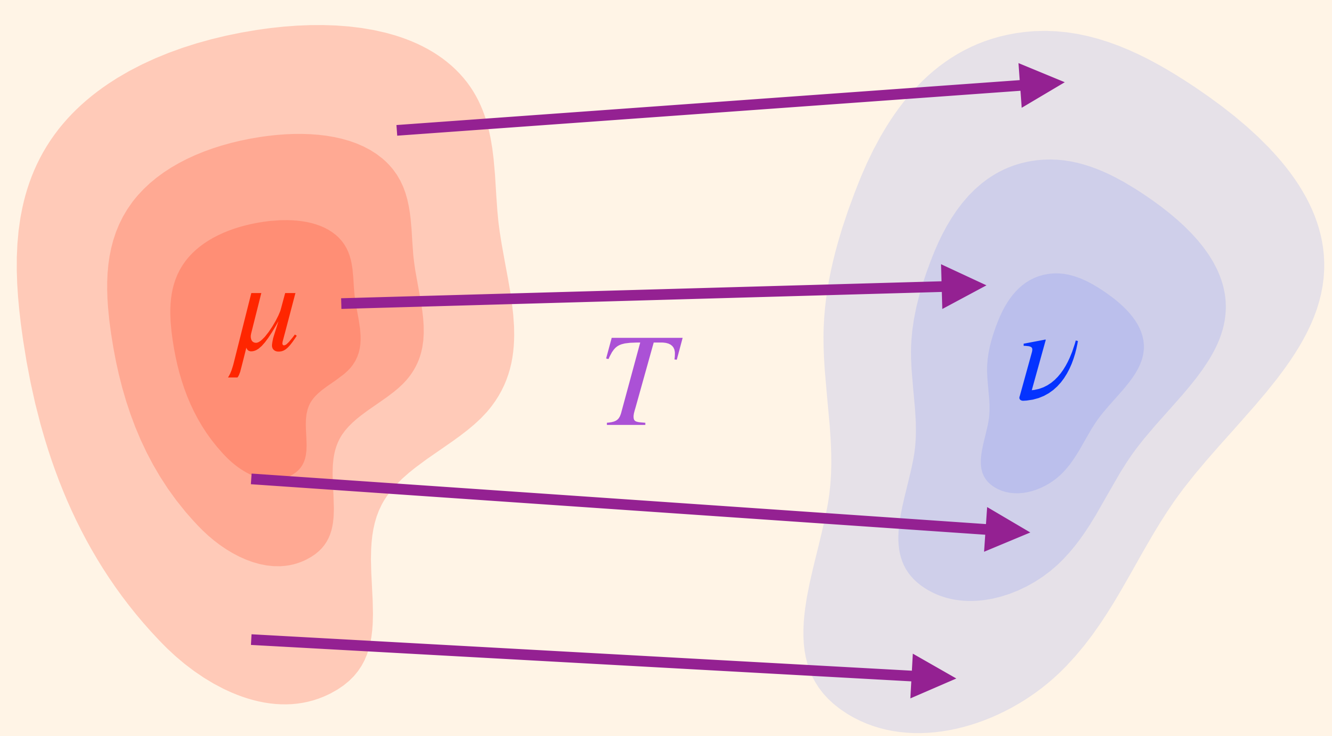


Monge 1784

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

Optimal Transport (Wasserstein) Distance





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

Monge 1784

Kantorovich 1942

General measures:

Kantorovich relaxation

or

Approximation by discrete measures

Universal Approximation

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

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] $\to H = 2$, dimension $\sim$ #tokens

[Geshkovski, Rigollet, Ruiz-Balet, 2024] $\to$ Universal interpolation.

[Agrachev, Letrouit 2019] $\to$ 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: $\gamma_\theta[\mu](x) := \langle x, u \rangle + \int \frac{e^{\langle Ax+b, y\rangle}}{\int e^{\langle Ax+b,y\rangle} \mathrm{d}\mu(y)} (\langle v, y \rangle + c) \mathrm{d}\mu(y)$ $\quad\theta := (A, b, c, u, v)$

→ First component of Attention ∘ MLP with skip connnexion.

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

Sketch of Proof

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

→ First component of Attention ∘ MLP with skip connnexion.

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

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.

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

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

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

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.

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

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

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

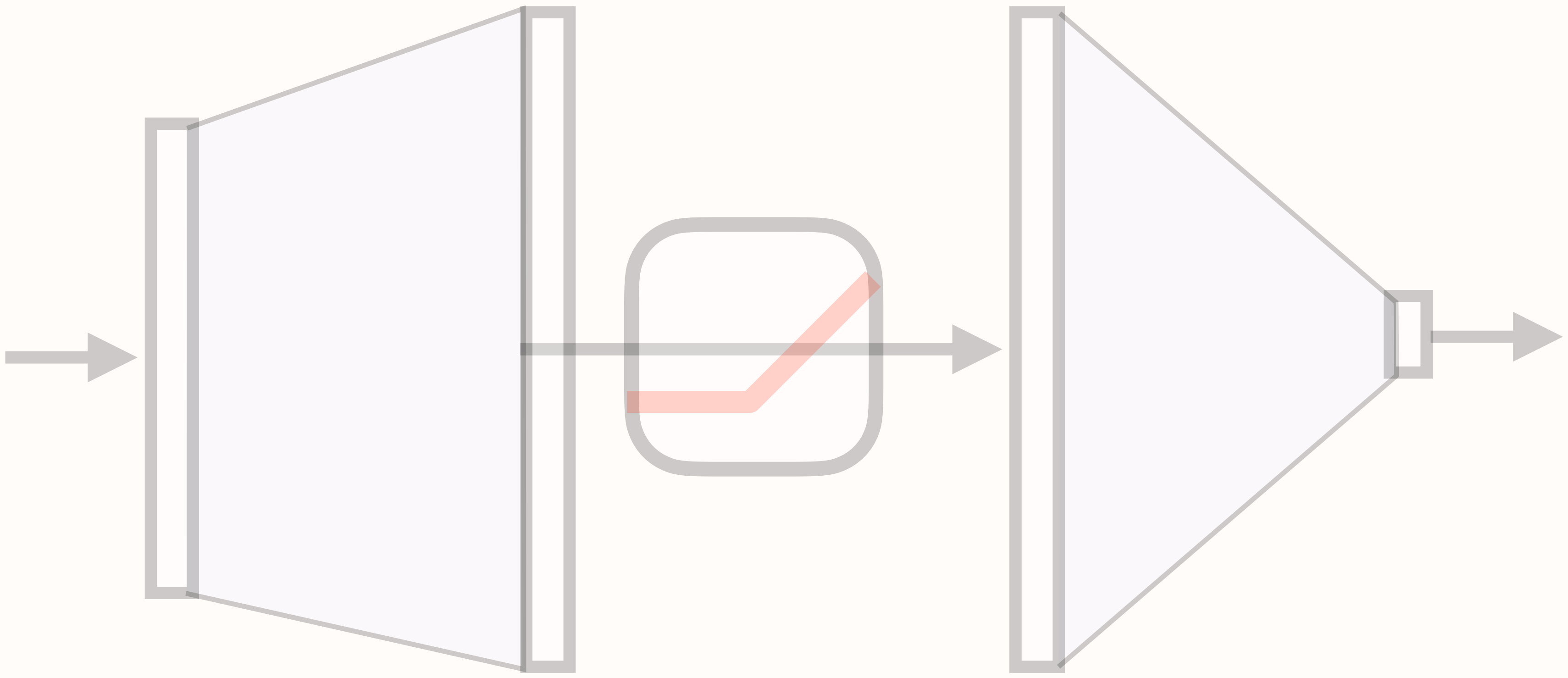
$A = c = u = 0$: $\quad L_1(\mu)(b) = L_1(\mu')(b)$

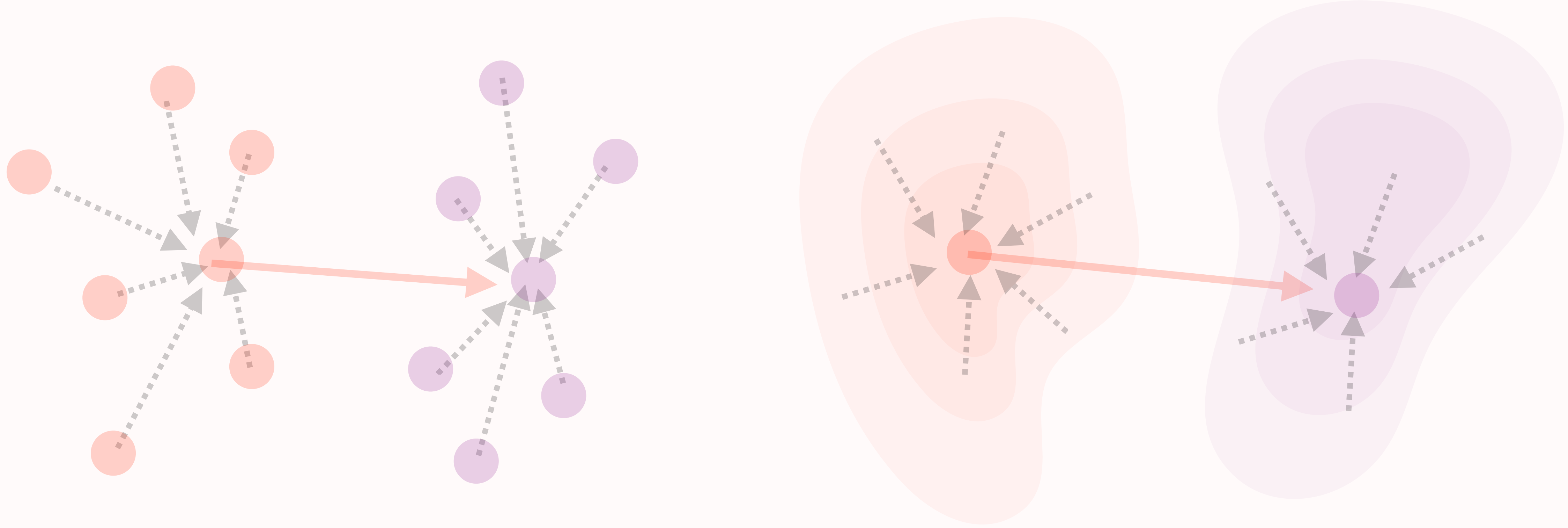
In 1-D: $\quad L_k(\mu)(b) := \int \frac{e^{by}y^k v}{\int e^{by'}\mathrm{d}\mu(y')}\mathrm{d}\mu(y)$

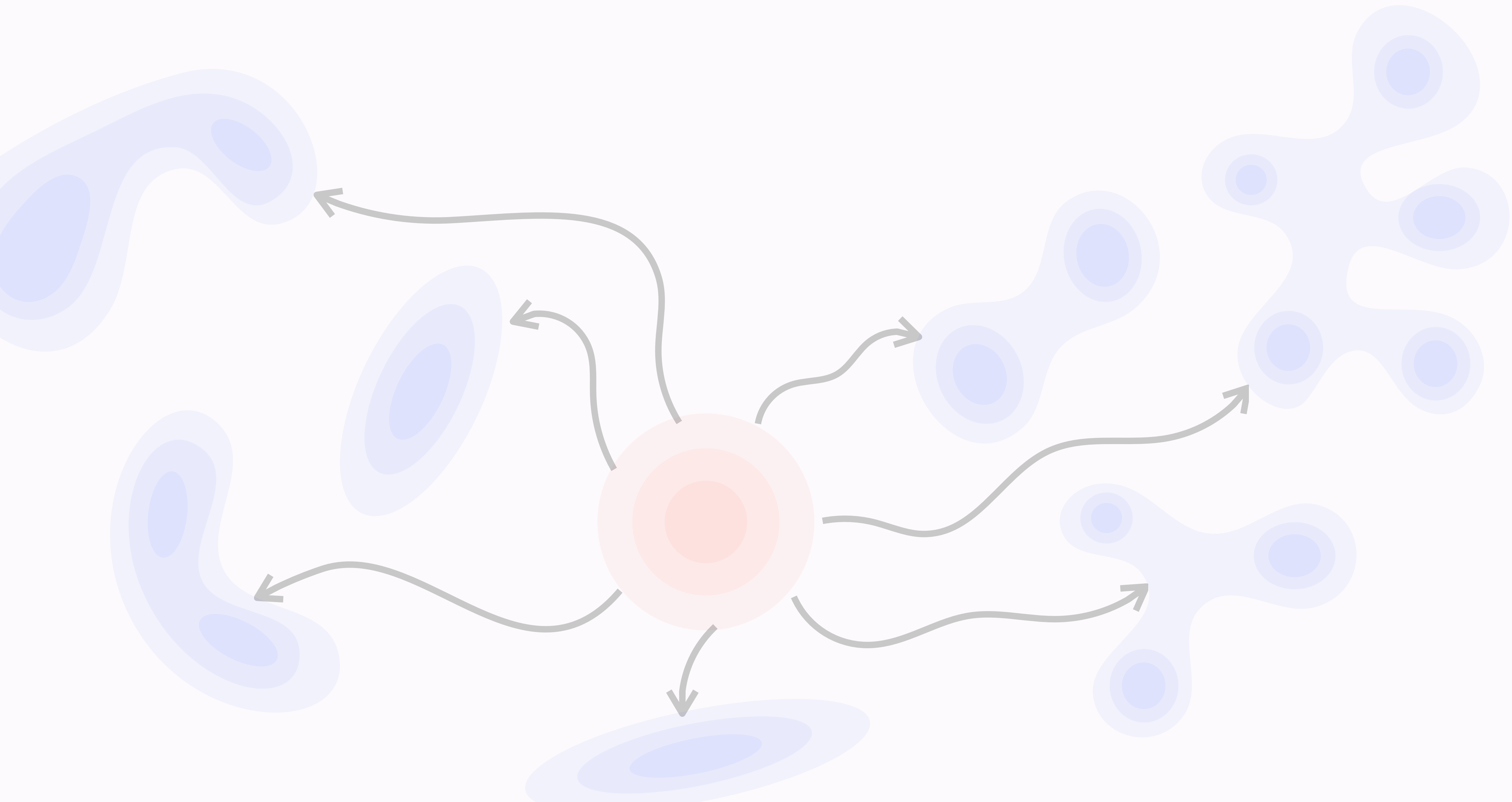
$$L_k' = L_{k+1} - L_k L_1$$

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

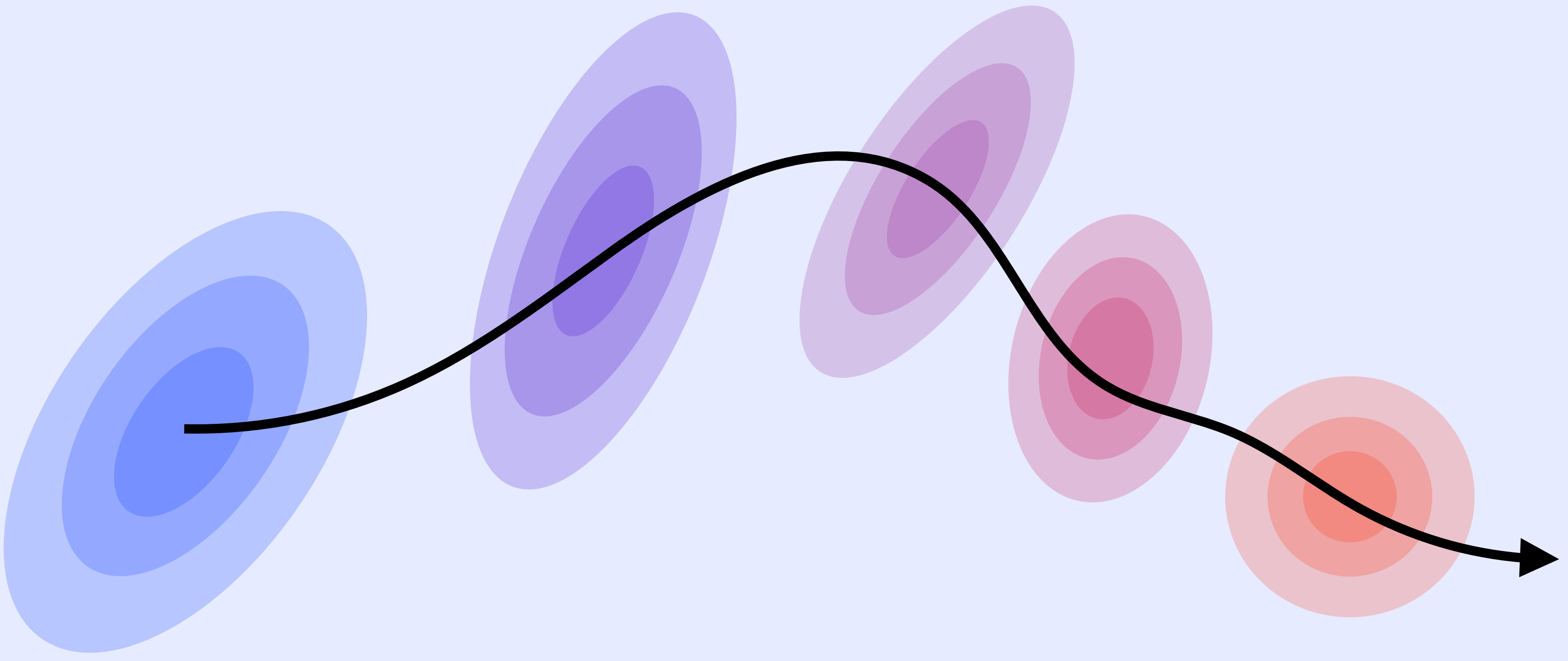
In higher dimensions: use Radon transform.







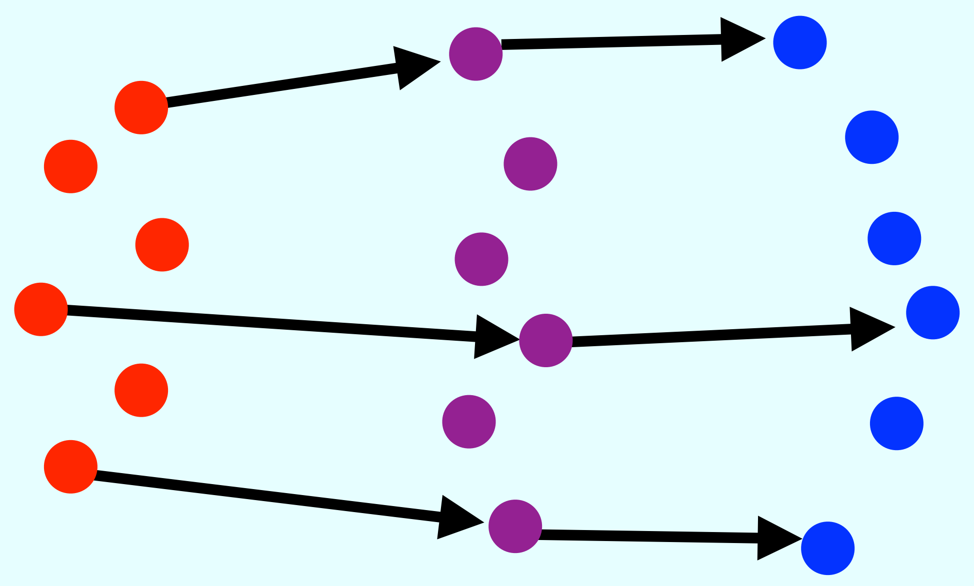
Infinite depth and PDEs



Infinite Depth as a Neural PDE

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

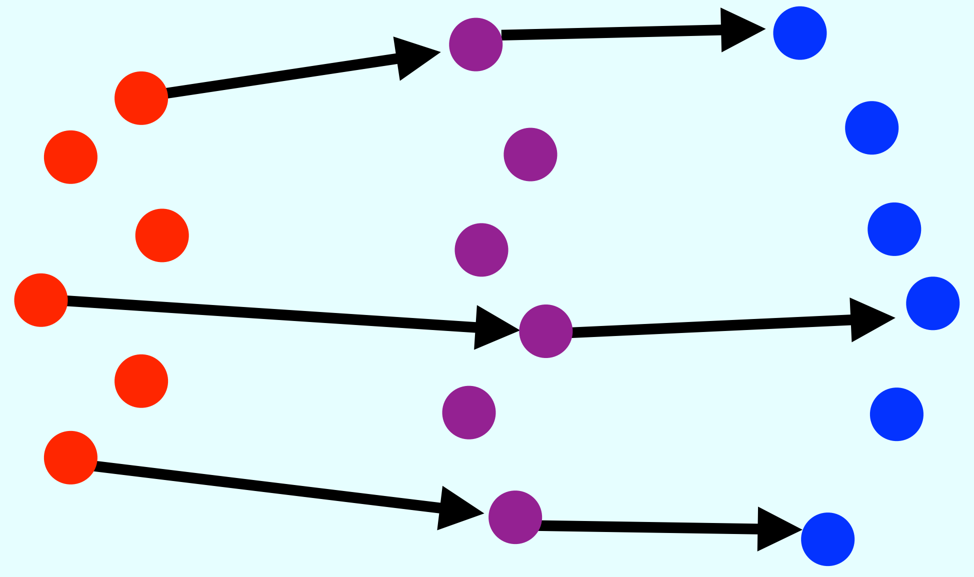
$$x_i(t+1) = x_i(t) + \frac{1}{T} \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$



Infinite Depth as a Neural PDE

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

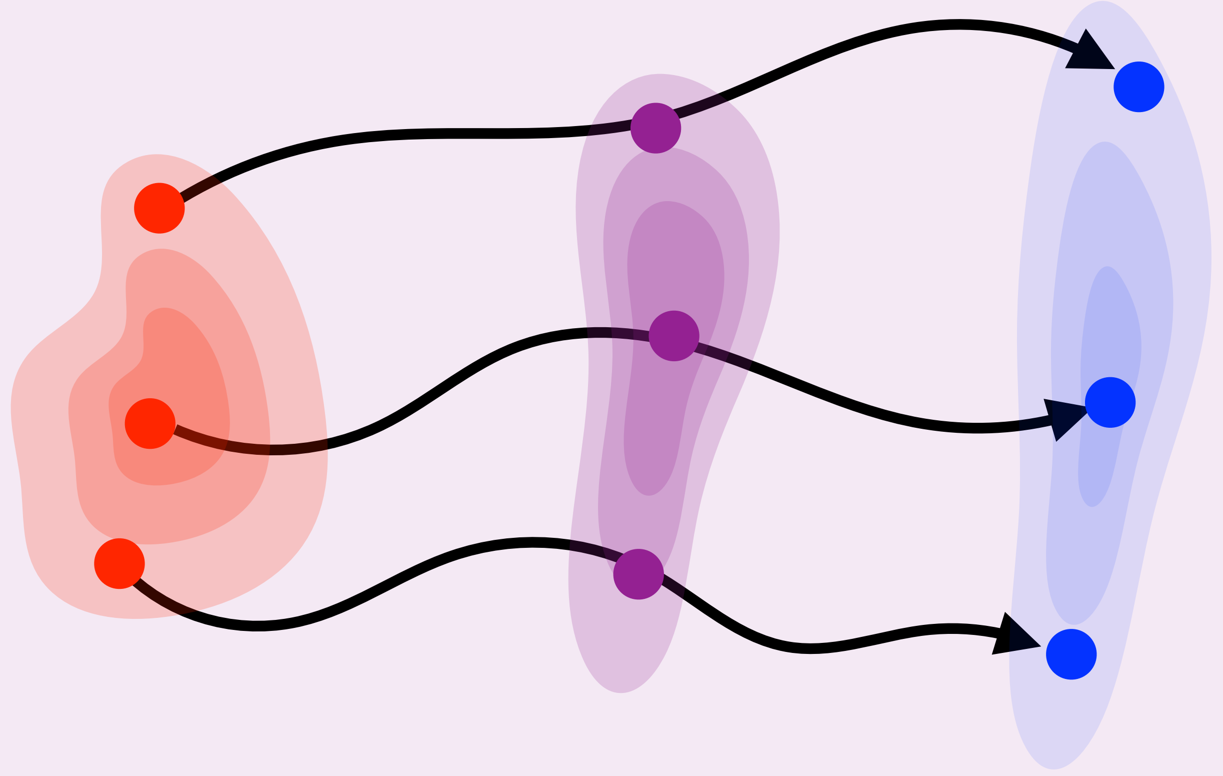
$$x_i(t+1) = x_i(t) + \frac{1}{T} \Gamma_{\theta(t)}[\mu(t)](x_i(t))$$



$$\xrightarrow[\text{Infinite depth}]{T \to +\infty}$$

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

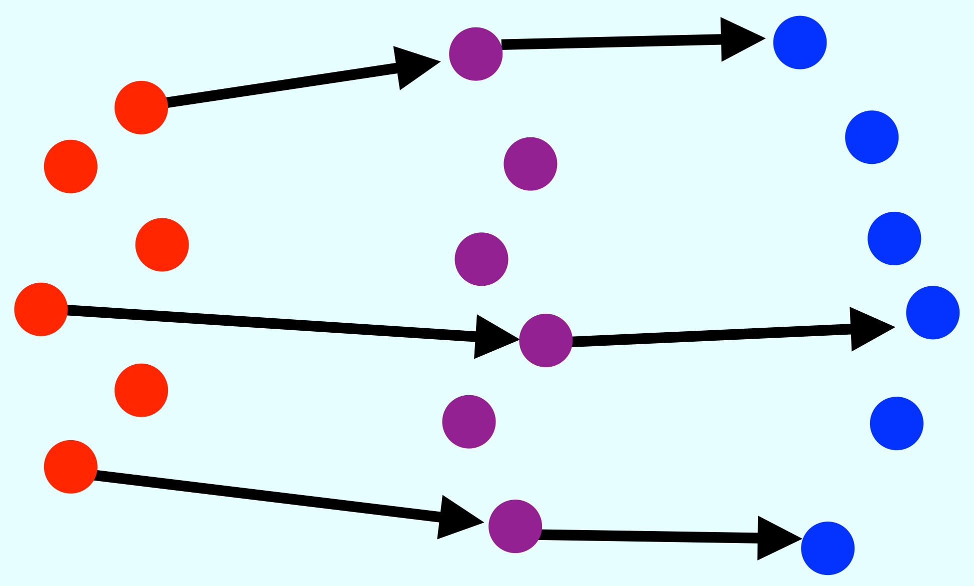
Coupled EDOs



Infinite Depth as a Neural PDE

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

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

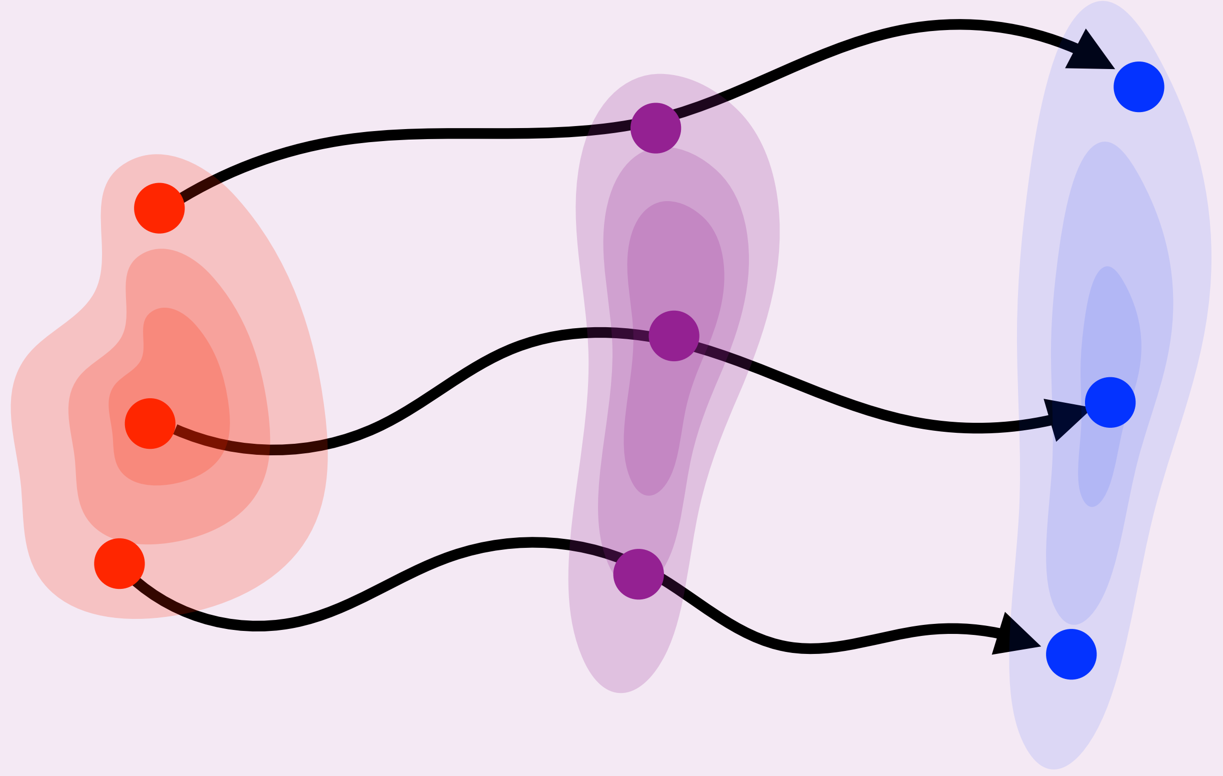


$T \to +\infty$

Infinite depth

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

Coupled EDOs



Mean field

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

Transformer PDE

→ Not a Wasserstein flow.

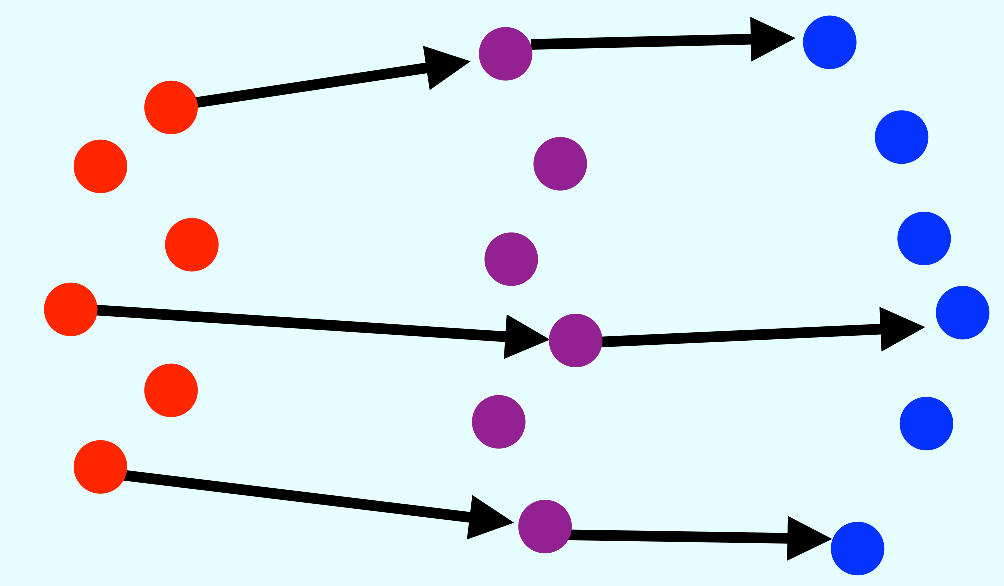
[Sander, Ablin, Blondel, Peyré, 2022]

[Geshkovski, Letrouit, Polyanskiy, Rigollet 2023]

Infinite Depth as a Neural PDE

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

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

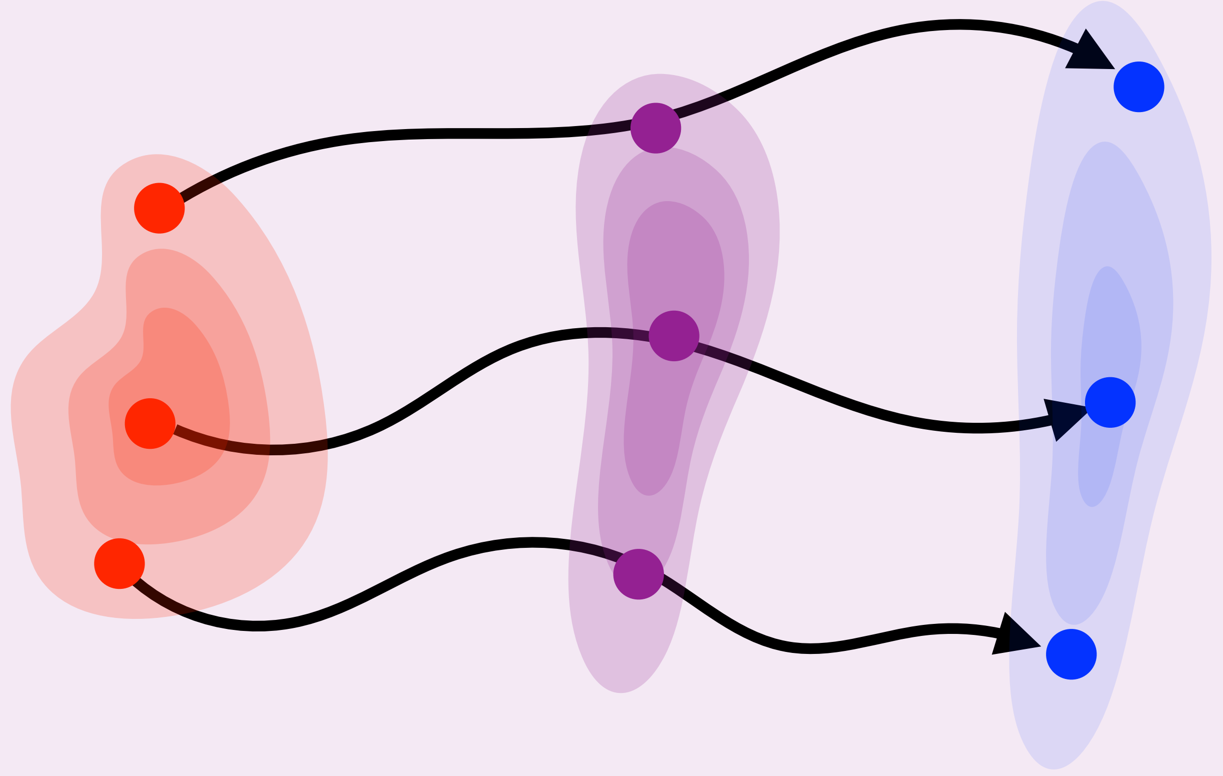


$T \to +\infty$

Infinite depth

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

Coupled EDOs



Mean field

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

Transformer 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

Gaussian Case and Clustering

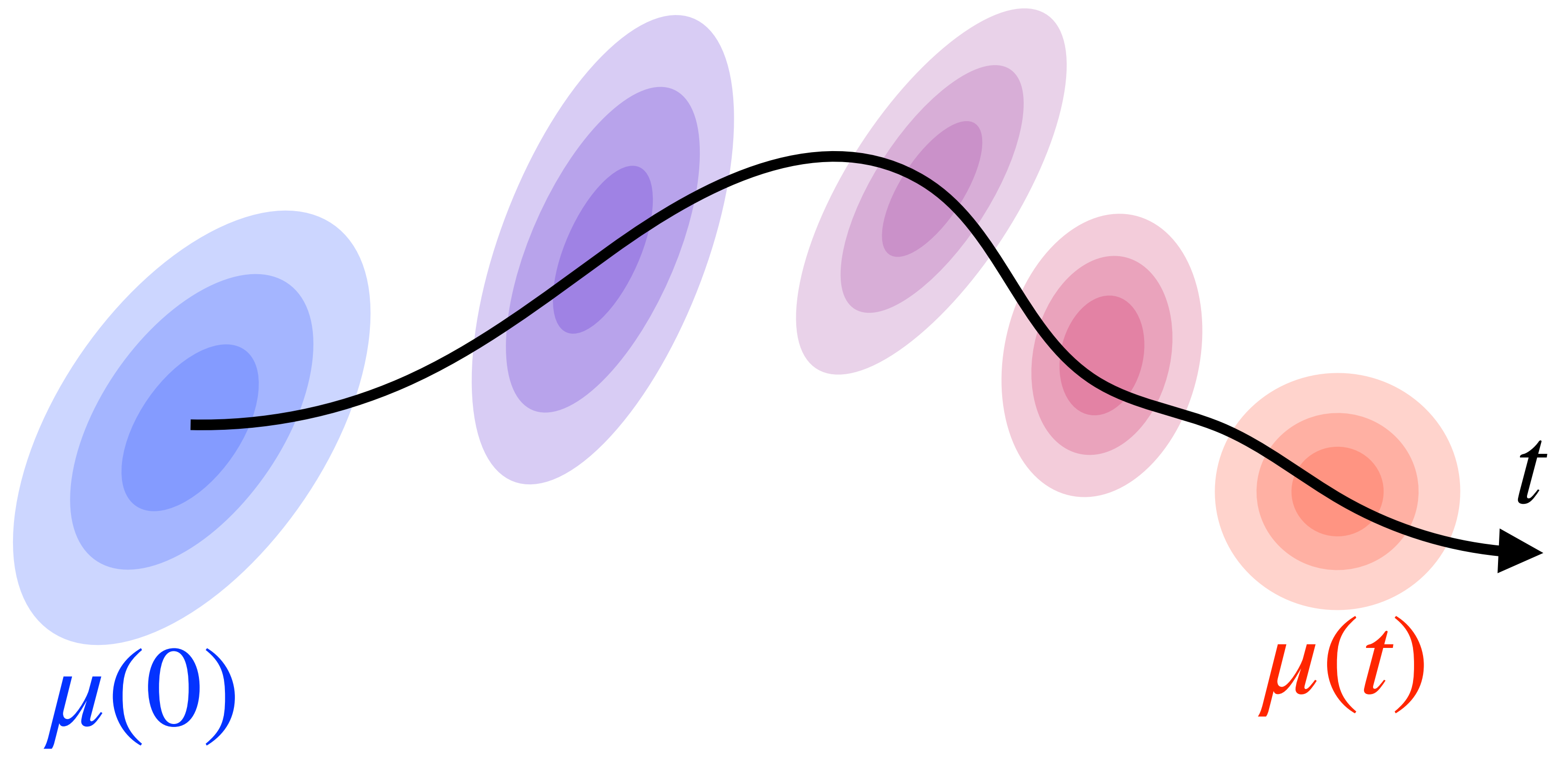
$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta(t) = (Q(t), K(t), V(t)) \qquad \frac{\mathrm{d}\mu}{\mathrm{d}t} + \mathrm{div}(\mu\Gamma_\theta[\mu]) = 0$$

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

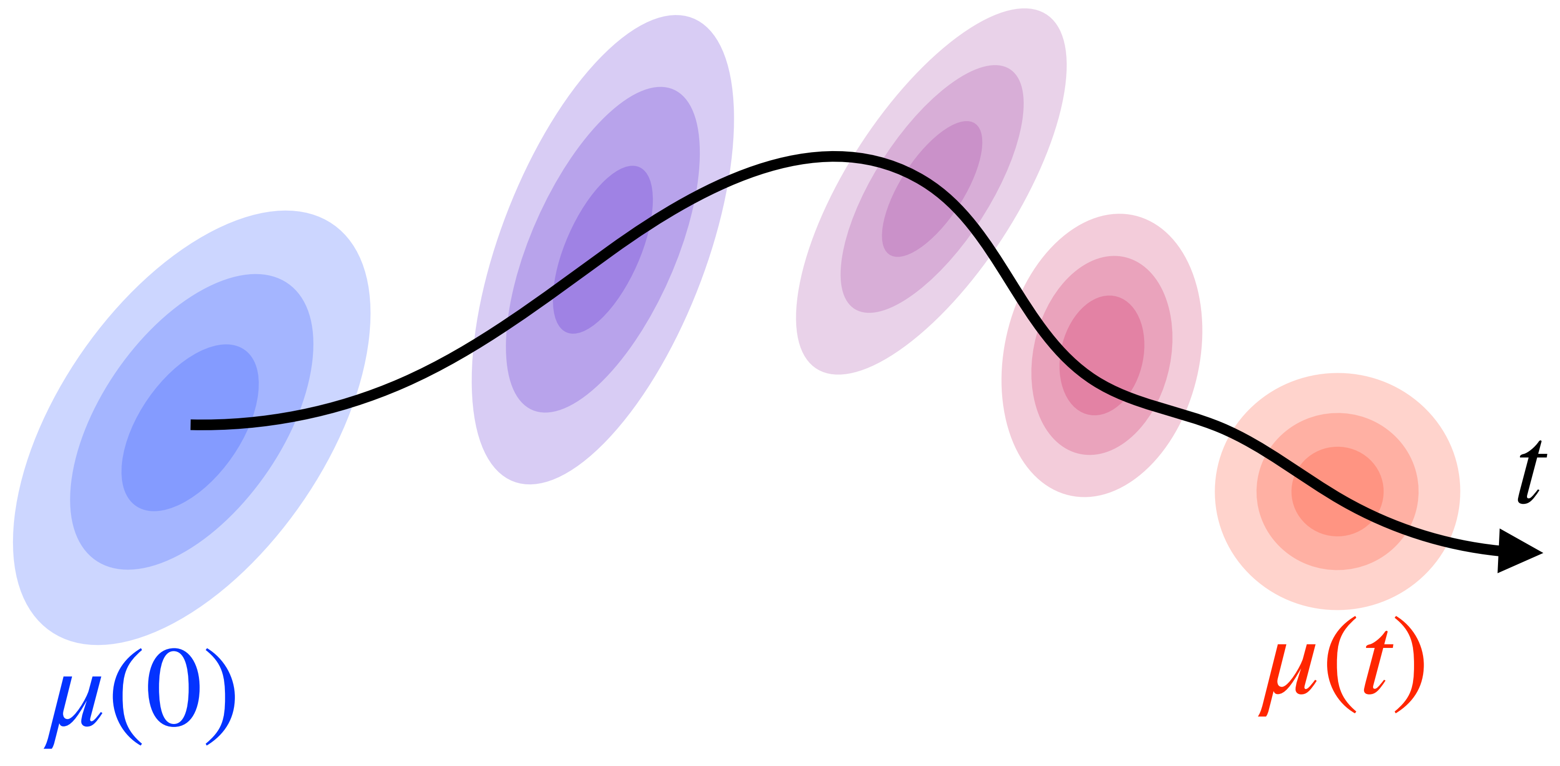
$$\Gamma_\theta[\mu](x) := \int \frac{e^{\langle Qx, Ky\rangle}}{\int e^{\langle Qx, Ky'\rangle} \mathrm{d}\mu(y')} Vy \, \mathrm{d}\mu(y) \qquad \theta(t) = (Q(t), K(t), V(t)) \qquad \frac{\mathrm{d}\mu}{\mathrm{d}t} + \mathrm{div}(\mu \Gamma_\theta[\mu]) = 0$$

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

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

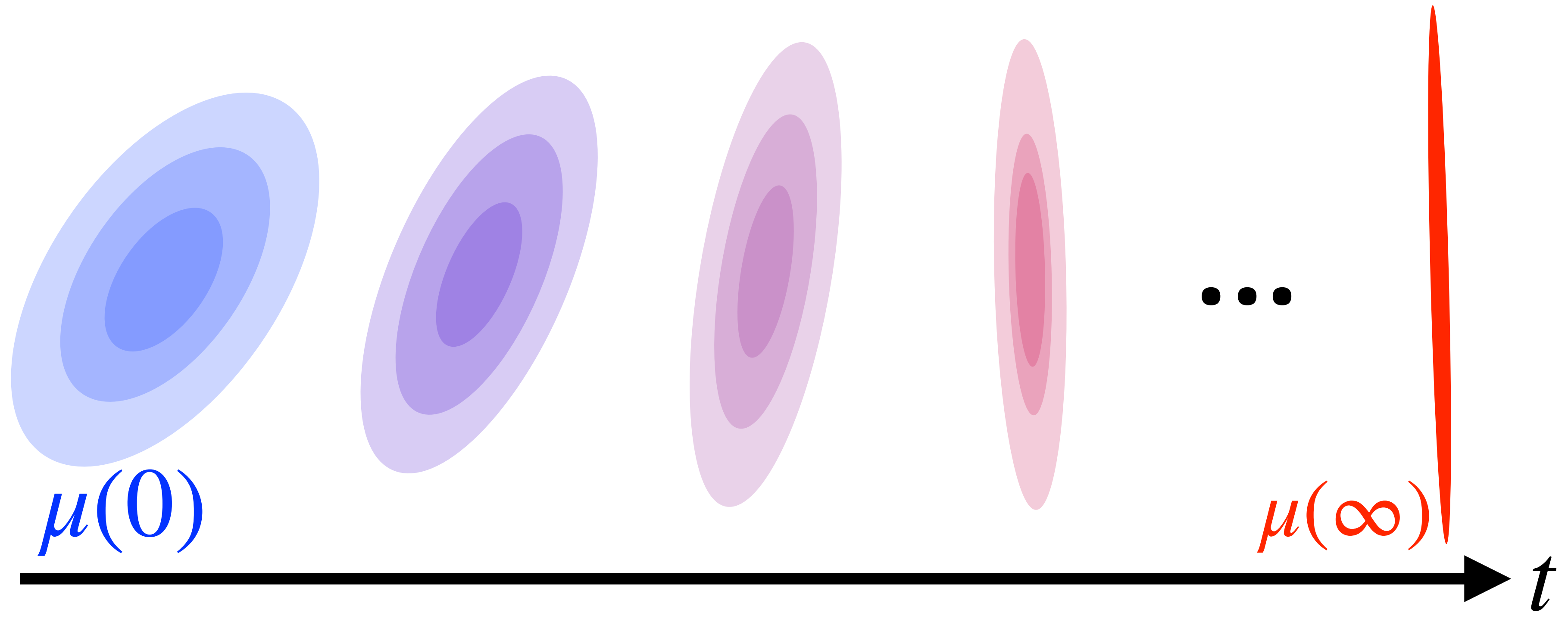
$$\dot{m} = V(\mathrm{Id} + \Sigma Q^\top K) m$$

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



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.

Open Problems and their Solutions

Universality: Toward quantitative approximation bound, leverage smoothness.

Optimisation: Understand the structure of optimal (Q, K, V)

Getting Albert interested in my problems